ABSTRACT

Title of dissertation: FANTASTIC SOURCES OF TUMOR HETEROGENEITY AND HOW TO CHARACTERIZE THEM

Sushant Patkar, Doctor of Philosophy, 2021

Dissertation directed by: Professor Eytan Ruppin
Department of Computer Science

Cancer constantly evolves to evade the host immune system and resist different treatments. As a consequence, we see a wide range of inter and intra-tumor heterogeneity. In this PhD thesis, we present a collection of computational methods that characterize this heterogeneity from diverse perspectives. First, we developed computational frameworks for predicting functional re-wiring events in cancer and imputing the functional effects of protein-protein interactions given genome-wide transcriptomics and genetic perturbation data. Second, we developed a computational framework to characterize intra-tumor genetic heterogeneity in melanoma from bulk sequencing data and study its effects on the host immune response and patient survival independently of the overall mutation burden. Third, we analyzed publicly available genome-wide copy number, expression and methylation data of distinct cancer types and their normal tissues of origin to systematically uncover factors driving the acquisition of cancer type-specific chromosomal aneuploidies. Lastly, we developed a new computational tool: CODEFACS (COnfident Deconvolution For All Cell Subsets) to dissect the cellular heterogeneity of each patient's

tumor microenvironment (TME) from bulk RNA sequencing data, and LIRICS (LIg- and Receptor Interactions between Cell Subsets): a supporting statistical framework to discover clinically relevant cellular immune crosstalk. Taken together, the methods presented in this thesis offer a way to study tumor heterogeneity in large patient cohorts using widely available bulk sequencing data and obtain new insights on tumor progression.

FANTASTIC SOURCES OF TUMOR HETEROGENEITY AND HOW TO CHARACTERIZE THEM

by

Sushant Patkar

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Eytan Ruppin, Co-Chair/Advisor
Dr. Soheil Feizi, Co-Chair
Professor Najib El-Sayed
Professor James A. Reggia
Dr. Furong Huang
Dr. Max Leiserson

Table of Contents

Preface

I completed my Bachelor's degree in 2016 in Computer Science at The Veermata Jijabai Technological Institute, Mumbai. I then joined the Computer Science PhD program at the University of Maryland (UMD), to pursue my passion for developing and utilizing techniques from the field of Artificial Intelligence (AI) and Machine Learning to solve real world problems. Soon after joining UMD, I started working in the interdisciplinary field of computational biology.

During my PhD, I had the good fortune to work with different labs, each with different research interests, hence the inherent "heterogeneity" of projects in this thesis. I learned a lot from these projects, but most importantly, I learned how regular interactions and collaborations with scientists from different disciplines help to comprehensively address the most challenging problems in biology and medicine. In this thesis, I discuss the overarching problem of heterogeneity in cancer and present computational techniques to characterize this heterogeneity from various perspectives and predict its clinical impact.

Dedication

I would like to dedicate this thesis to all my friends and family for their unconditional support throughout this journey and my aunt who passed away from pancreatic cancer.

Acknowledgments

This body of work would not have been possible without the help and support of several people.

First and foremost, I would like to express my profound gratitude and appreciation to my current advisor, Dr. Eytan Ruppin, for his continuous guidance and support, especially since the COVID-19 pandemic hit the world. Thank you for teachning me how to ask the right research questions, how to effectively present one's work, and most importantly how to keep one's spirits lifted in times of uncertainty.

Second, I would like to thank my previous advisors Dr. Sridhar Hannenhalli and Dr. Roded Sharan, who introduced me to the field of Computational Biology and taught me how to effectively navigate collaborative research projects.

Third, I would like to thank Dr. Rajiv Gandhi who introduced me to the field of Computer Science and Algorithms. Thank you for helping me cultivate the necessary discipline and attitude to pursue a career in research.

I would also like to thank all current and former members of the different labs with whom I've had the good fortune to work and learn from: Dr. Alejandro A. Schaffer, Dr. E. Mike Gertz, Dr. Fiorella Schischlik, Dr. David Crawford, Welles Robinson, Dr. Noam Auslander, Dr. Erez Persi, Rotem Katzir, Dr. Joo Sang Lee, Dr. Kevin Litchfield, Dr. Assaf Magen, Dr. Kun Wang, Dr. Mahfuza Sharmin and Dr. Dana Silverbush. A special thanks to Dr. Noam Auslander for her indispensable guidance in the second half of my PhD.

Lastly, I would like to thank all the excellent experimental collaborators with

whom I was fortunate to work: Dr. Yardena Samuels, Dr. Osnat Bartok, Dr. Yochai Wolf, Dr. Soma Ghosh, Dr. Yossi Yarden and Dr. Thomas Reid.

It is impossible to remember all, and I apologize to those I've inadvertently left out. Lastly, thank you God!

List of Tables

List of Figures

List of Abbreviations

ILP	Integer Linear Program
KPI	Kinase Phosphatase Interactions
PDI	Protein DNA Interactions
ASP	A-Shortest Path
AllSP	All-Shortest Paths
AdirSP	A directed Sortest Path
TCGA	The Cancer Genome Atlas
KEGG	Kyoto Encyclopedia of Genes and Genomes
AUC	Area Under the Receiver Operator Characteristic Curve
ITH	Intra Tumor Heterogeneity
VAF	Variant Allele Frequency
SCC	Single Cell Clone
UVB	Ultra Violet B light
TMB	Tumor Mutation Burden
CNV	Copy Number Variation
SDI	Shannon Diversity Index
GTEx	Genotype-Tissue Expression
LAML	Acute Myeloid Leukemia
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma

SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine Carcinosarcoma
UCEC	Uterine Corpus Endometrial Carcinoma
UVM	Uveal Melanoma
TME	Tumor Micro Environment
CODEFACS	Confident Deconvolution For All Cell Subsets
LIRICS	Ligand Receptor Interactions Between Cell Subsets
MSI	Microsatellite Instability
MMRD	Mismatch Repair Deficient
MSFI	Mutation Specific Functional Interactions
PFS	Progression Free Survival
OS	Overall Survival
DC	Dendritic Cell
CAF	Cancer Associated Fibroblasts
NK	Natural Killer
KM	Kaplan Meier

Chapter 1

Introduction

1.1 The problem of heterogeneity in cancer

Cancer is a heterogeneous disease. This heterogeneity is fueled by evolution and eventually leads to resistance to chemo and targeted therapies [139]. Although the recent success of immunotherapy has brought us closer to a cure, a large fraction of patients still fail to receive durable clinical benefit [216]. In order to stay one step ahead in this evolutionary arms race, it has become clear that one needs to design rational treatment combinations. Key to designing such combinations is our ability to characterize the extent of tumor heterogeneity, both among and within patients, and identify the underlying evolutionary mechanisms driving it.

Projects like The Cancer Genome Atlas [1], vastly improved our knowledge on the extent of tumor heterogeneity among patients by doing a comprehensive multi-"omic" profiling of thousands of patient tumors (inter-tumor heterogeneity). This revealed novel tumor subtypes and their underlying driver mutations. With recent advances in next generation sequencing technologies, it is now possible to further sequence tumors at a single cell resolution, thereby allowing one to assess the heterogeneity of tumors within patients (intra-tumor heterogeneity).

While single cell data ideally represents the state-of-the-art in terms of characterizing tumor heterogeneity, in this thesis we predominantly focus on developing

computational methods to characterize tumor heterogeneity from bulk sequencing data and demonstrate their clinical impact; something which is currently not possible using single cell technologies alone due to their limited scalability to large cohorts.

1.2 Fantastic sources of tumor heterogeneity and ways to characterize them

There exist multiple sources of tumor heterogeneity. However, they can be broadly categorized into the following three groups:

1.2.1 Genetic Heterogeneity

Cancer is, in essence, a genetic disease [237]. Errors in the DNA replication or repair machinery result in accumulation of somatic mutations in cells, and in specific contexts, these mutations can drive multiple clonal expansions [144]. This gives rise to heterogeneous tumor cell populations. In addition, a vast majority of tumors have chromosomal instability, which results in large-scale structural changes (such as the gain or loss of entire chromosomes), that are often associated with poor clinical outcomes[22, 226, 56]. In chapter 3, we show how computational methods can be applied to characterize intra-tumor genetic heterogeneity from bulk sequencing data and study its effects on the host immune response while controlling for differences in overall mutation burden. In addition, in chapter 4, we investigate chromosome arm imbalances across solid tumors from different tissues of origin to uncover factors

explaining their observed tissue-specific heterogeneity.

1.2.2 Epigenetic Heterogeneity

Tumor cells can also exhibit phenotypic plasticity by dynamically modulating gene expression via epigenetic mechanisms or cellular signalling events. This is often said to lead to acquired resistance to many targeted therapies as opposed to hardwired resistance that arises from genetic heterogeneity [139]. In chapter 2, utilizing bulk transcriptomic data of breast cancer patients from the TCGA, we explore how such dynamic changes in gene expression in cancer can potentially lead to a "functional re-wiring" of genes. Such functional re-wiring could explain why a vast majority genetic interactions are context specific, making the translation of these interactions into clinically effective targeted therapies a challenge [11]. In addition we also developed new mixed integer linear programming frameworks to impute functional effects of protein-protein interactions from genetic perturbation data, thereby bringing us one step closer to building accurate computational models of cellular signaling [157].

1.2.3 Microenvironmental Heterogeneity

Tumor cells also constantly interact with other cell types in their microenvironment in order to facilitate their growth and suppress the host immune response. These interactions can further fuel tumor evolution and lead to their observed genetic and epigenetic heterogeneity. It has become increasingly evident that

the interactions between cell types in the tumor microenvironemt play a critical role in facilitating a response or resistance to treatment with the establishment of immunotherapy as the third major arm of cancer treatment (besides surgery and chemo therapies) [216]. In chapter 5, we develop a new computational method to characterize the cellular heterogeneity of each patient's tumor microenvironment from bulk transcriptomic data. Using our method one can aim to not only infer the cellular abundance of each cell type in the tumor micro-environment but also investigate their transcriptional states and infer clinically relevant cellular crosstalk. Applying our method to the TCGA, we generate a large resource of deconvolved transcriptomes of each patient's tumor tumor sample, thereby enabling the analysis of the TCGA at a cell type specific resolution. In addition, we uncover a shared repertoire of cell-cell interactions that specifically occur in the TME of mismatch-repair-deficient solid tumors and explain their universally high response rates to immune checkpoint blockade treatment. These results point to specific T-cell co-stimulating interactions that can enhance immunotherapy responses in tumors independent of tumor mutation burden levels. Finally, using machine learning, we demonstrate how one can exploit the large deconvolved data resource we generated to identify key cell-cell interactions in the TME predicting patient responses to immune checkpoint blockade therapy in melanoma.

Chapter 2

Algorithms for context-specific functional annotation of genes and the imputation of functional effects of protein-protein interactions

⋆⋆ This work was done in collaboration with Dr. Sridhar Hannenhalli and Dr. Roded Sharan and appears in PLoS Computational Biology [171] and Bioinformatics [172]

2.1 Overview

Cellular functions are carried out by networks of interacting proteins [17]. In particular, empirical data suggest that proteins that participate in the same biological process or a pathway tend to interact with one another, and more broadly, tend to inhabit the same neighborhood in the protein interaction network (PIN). This guilt-by-association principle has been successfully applied to predict protein function, outperforming alternative methods that do not take the PIN into account [5, 123, 138, 206, 207, 223].

Given that a gene's function is informed by its PIN neighborhood, it is plausible that a an organism may dynamically adapt its genes' functions across different contexts, such as developmental stages, tissues, diseases and evolution, by altering the PIN structure. For example, during Drosophila development, a key regulatory transcription factor fushi tarazu (FTZ) changes function from an ancestral homeotic

gene (those that regulate development of specific body parts) to a pair-rule segmentation gene (regulating initial formation of the segments in a developing embryo). Notably, this functional switch involves changes in FTZ's interaction partners; while in the ancestral species FTZ interacted with homeotic proteins, in drosophila it interacts with protein involved in segmentation, and thus it got co-opted into segmentation function [134]. Furthermore, many genetic interactions exhibit context specificity [96]. Based on these premises, we describe a network diffusion-based algorithm to predict how a gene's function might change due to shifts in its protein-protein interaction neighborhood during malignant transformation. This approach uniquely reveals several functions that are significantly lost or gained in breast cancer and modulate patient survival.

Furthermore, another aspect important for the characterization epigenetic tumor heterogeneity is modeling how cells respond to different genetic alterations or environmental perturbations. Key to building such models is having well annotated biological pathways representing how biological signals flow affect the activity of different proteins and transcription of genes. However, such annotations are sparse. So we additionally developed an optimization framework to impute missing functional annotations describing how biological signals are propagated over the joint protein-protein interaction and regulatory network of a cell given transcriptional data before and after genetic perturbations. This imputation problem was previously shown to be non-deterministic polynomial time (NP)-hard for general networks [25]. In this work, we overcome the limitation of network coverage of previous methods by developing new mixed integer linear programming formulations and utilizing

state-of-the-art SAT solvers. Overall, our imputation method outperforms previous work by a considerable margin.

2.2 Methods

2.2.1 Inference of gain or loss of function via "guilt-by-association"

Let $G(V, E)$ be the weighted un-directed network with V representing the set of nodes and E the set of weighted interactions. Let W be the weighted adjacency matrix corresponding to G and let D be the diagonal degree matrix (with diagonal entries corresponding to the weighted degree of each node in the graph). For a biological function f (which represents any biological process or pathway in a publicly available database), let A_f be the set of genes annotated with that function. For a RNA-seq sample s, let $G_s(V_s, E_s)$ be the sample-specific sub-network of G consisting of all genes with an expression ≥ 1 RPKM in that sample, let Y_s be the prior knowledge vector such that $Y_{s,g} = 1, \forall g \in A_f \cap V_s$. The guilt-by-association principle implies that the involvement of any gene g in a function f is likely to be influenced by the involvement of the genes in its neighborhood (Figure 2.1). Additionally, the involvement should be consistent with our prior knowledge of functional memberships. This can be mathematically modelled by the following diffusion equation:

$$F_s = (1 - \alpha)(D_s^{-1/2} W_s D_s^{-1/2}) F_s + \alpha Y_s \tag{2.1}$$

Here F_s is a vector of raw involvement scores of every gene in G_s.$\alpha \in (0, 1)$ is a parameter that weighs the importance of prior knowledge in the model. Notice that the adjacency matrix W_s is symmetrically normalized by the square root of the product of node degrees. This step controls for biases that may arise from diffusing information through high degree nodes (hubs) in the network. As shown in previous work, the raw scores are fairly robust for the choice of α, and we adopt the choice $\alpha = 0.2$ following [234]. Since the Eigenvalues of $D_s^{-1/2} W_s D_s^{-1/2}$ lie in $[-1, 1]$, it can be shown that $I - (1-\alpha) D_s^{-1/2} W_s D_s^{-1/2}$ is positive definite and we get the following solution:

$$F_s = \alpha (I - (1 - \alpha) D_s^{-1/2} W_s D_s^{-1/2})^{-1} Y_s \tag{2.2}$$

There are several ways to compute the above solution, the simplest being the iterative matrix multiplication algorithm first proposed by Zhou [34]. To circumvent the overhead costs of multiplying large matrices, we proceed by solving the system using the conjugate gradient (CG) method. The above procedure assigns a raw involvement score to each gene in G_s for each diffused function. This raw score however depends on$|A_f \cap V_s|$ as well as the sample-specific PIN topology. To appropriately calibrate it, we can estimate a significance p-value for the score, in a function-specific manner. This is done by comparing a gene's raw score against a null distribution of scores generated by diffusing random prior knowledge vectors in G_s annotating $|A_f \cap V_s|$ genes. Hence each null distribution is parameterized by $|A_f \cap V_s|$ which we call the seed size. Note that this technique requires us to

run a large number of bootstrap instances on separately for each sample-specific PIN (1157 samples in total analyzed in our study) for each function (1184 functions evaluated in this study). To tackle such an enormous computational task, we follow a memoization procedure in which for each sample, we pre-compute a smaller set of null distributions from pre-determined seed sizes (40 to 500 with an increment of 10) and estimate p-values of diffused raw scores by simply comparing them to a null distribution closest in seed size to the true null distribution for that sample. The null distributions are based on 100 bootstrap samples. Finally, we say that a gene is assigned a function f in a given sample if the p-value associated with its raw score in that sample < 0.01.

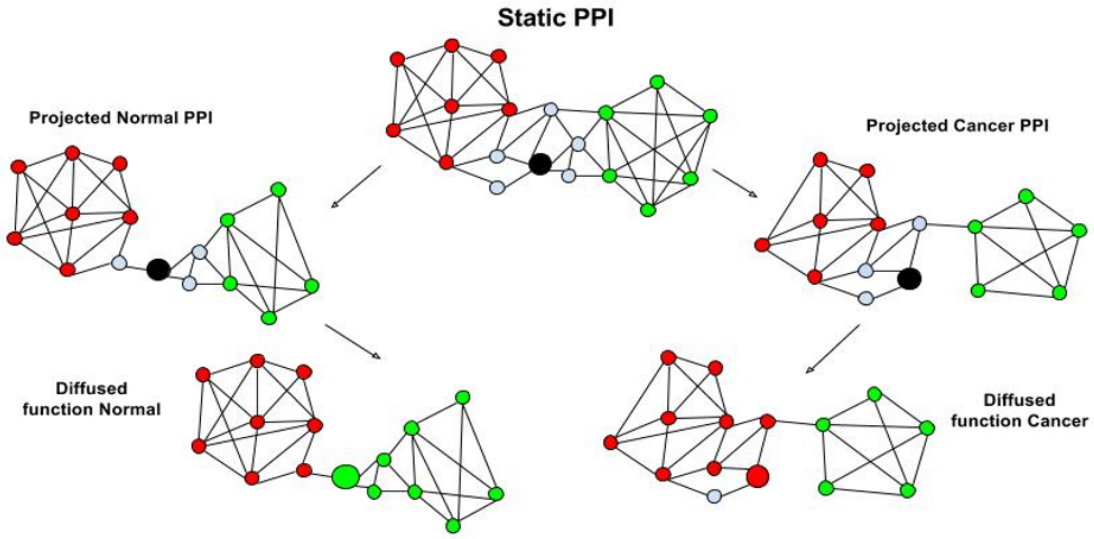

Figure 2.1: The reference gene is depicted by black circle. The initial static global PIN is projected onto normal and cancer samples based on gene expression, and each function (red and green) are diffused through each PIN. In this case, the reference gene is assigned green function in normal and red function in cancer, i.e., the gene gained red and lost the green function in cancer.

Given a cohort of samples under two conditions (normal and breast cancer in our application) and a gene-function pair (g, f), we determine the number of sam-

ples where g was gained function f by diffusion (see above). Having determined this separately for normal and breast cancer samples, we perform a Fisher's exact test to assess whether the gain of f by g is significantly enriched in either one of the conditions. We say a gene g gains a function f in cancer if the gain of f by g is significantly enriched among cancer samples when compared to normal. The enrichment p value is determined by Fisher exact test. Unless stated otherwise, we use the default p-value significance threshold of 0.05. We also estimated the False Discovery Rate (FDR) for each pair of g and f. The FDR criterion however yields substantially fewer genes resulting in decreased power for various downstream analyses. Therefore by default we used the p-value criterion, and provide the results based on FDR criterion in the supplementary material of the publication. In addition to the significance criteria, we also consider the effect size of the functional gain or loss. Let θ be the odds ratio derived from the Fisher contingency table. To reduce chances of false discovery, we require the effect size to be large. Hence, for downstream analyses we looked at a range of θ from $\theta = 2$ to 10, and unless otherwise mentioned, the default is highly stringent $\theta = 10$, while the results for other values of θ are provided in the supplementary material of the publication. Note that if a gene is not expressed in a sample then it is not present in the sample-specific PIN and therefore cannot be assigned a function. Thus if g is un-annotated by a function, biases may arise in the determination of its gain or loss via guilt by association if there are significant differences in the expression of g in sample-specific networks generated within a cohort or between two cohorts. To control for such a bias, we take two filtering measures. First, we check if g is expressed significantly

more in samples corresponding to one condition relative to the other by building a contingency table for expressed versus not expressed among normal and cancer samples and performing a Fisher exact test. We exclude g if its p-value ≤ 0.05. Second, in estimating loss and gain for g relative to a function we only consider samples where g was expressed. This results in downstream analyses of 12599 genes out of a total of 16562 from the original network. To quantify the degree of gain (or loss) of a function in cancer relative to normal due to guilt by association, we only consider the genes that are not annotated to have that function. This ensures that our estimated change in functional activity is informed primarily by the changes in PIN topology and not by the differential expression of the genes annotated to perform a certain function. We Define

$$\Phi(f,g) = \begin{cases} 1 & \text{if } g \text{ is un-annotated and gains } f \text{ with } \theta \geq 10 \\ -1 & \text{if } g \text{ is un-annotated and loses } f \text{ with } \theta \leq 0.1 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Let $\Delta_f = \sum_g \Phi(f,g)$ be the difference between the number of un-annotated genes gaining and losing f. The higher the $|\Delta_f|$ value, the greater the change in activity of f between normal and cancer. The direction of change is determined by the sign: "+" represents increase in activity from normal to cancer due to a greater number of un-annotated genes potentially acquiring that function in cancer; we refer to such a function as *cancer-associated gained function*. Likewise, "-" represents an

overall decrease in functional activity due to a greater number of un-annotated genes potentially losing that function in cancer; we refer to such a function as *cancer-associated lost function*.

2.2.2 A mixed integer linear programming (ILP) framework for imputation of functional effects of protein-protein interactions

In this section we describe algorithms for inferring signs and direction of signal flow over the protein-protein/protein-DNA interaction network. The sign represents the functional effect on the target gene/protein carrying the signal along the network. This depends on the type of the physical interaction being considered. For protein-DNA interactions (PDIs), a $+/-$ sign describes a regulatory effect; for protein-protein interactions (such as phosphorylation/de-phosphorylation interactions between kinases and phosphatases), it represents a functional activation/repression effect. Currently, such direction and sign information is available for only a few well-studied pathways (see Figure 2.2 for an example), although a large fraction (40-70%) of the PPIs are expected to admit such an annotation [213]. The inference of such annotation information is a precondition to any logical model of a system under study (see, e.g., [157]). We start by formally defining the problem and sketching the previous approach of [97]. Then, we study three variants of the original problem (each describing a signaling model) and develop novel integer linear programming formulations to solve them to optimality on current networks. We assume we are given a (potentially partially signed) physical interaction network along

with a collection of cause-effect gene pairs, such as commonly obtained from knock-out experiments. The Maximum sign assignment (MSA) problem is to assign signs to the unsigned edges of the network in a way that best explains the given pairs. We say that a cause-effect pair (s, t) with sign δ_{st} ($+$ encoding down-regulation of t in response to the knockout of s, $-$ encoding up-regulation of t in response to the knockout of s) is *explained* or *satisfied* by a sign assignment, if there exists a path in the network from s to t whose aggregate sign (the product of the signs along its edges) is δ_{st}. Formally, MSA is defined as follows:

Input. *A partially signed network* $G(V, E)$ *and a set of* k *cause-effect pairs* $(s_1, t_1), \ldots, (s_k, t_k)$ *with signs* $\delta_{s_1 t_1}, \ldots, \delta_{s_k, t_k} \in \{+, -\}$

Goal. *A sign assignment to the unsigned edges of the network such that a maximum number of input pairs are satisfied by the assignment.*

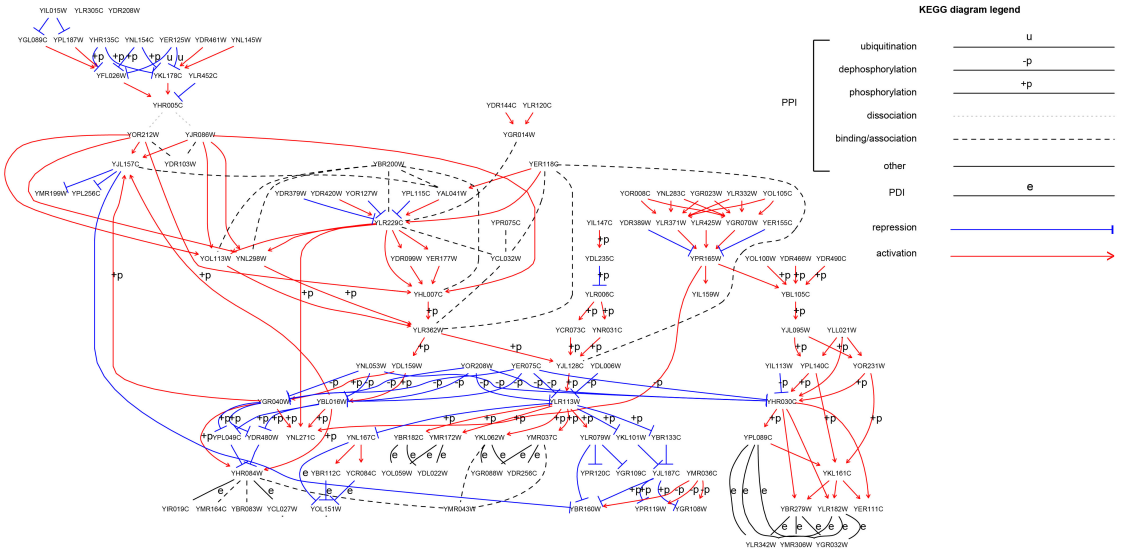

Figure 2.2: Yeast signaling pathways from KEGG in one network depicting the organization of different types of physical interactions with their respective experimentally-derived annotations of signal flow.

This problem focuses on the A-path signaling model of [253]. [97] showed that due to the nature of the model, any unsigned edge that lies on a cycle in the network cannot be uniquely signed. They generalized this notion to any 2-connected component (or block) by determining if these components are *strongly signed*. They then proposed an approach to reduce the input network to an acyclic one by contracting all edges in these strongly signed components without affecting the maximum number of pairs that could be satisfied. In the reduced network, every pair is connected by a unique path, facilitating the formulation of an ILP to assign signs to the unsigned edges of this path such that the number of satisfied pairs is maximized. A key drawback of this approach is that reducing the network to an acyclic one severely restricts the number of edges participating in the ILP (coverage) and, hence, restricts the number of interactions that can be uniquely signed. In subsequent paragraphs, we discuss three variants of MSA, each describing a different plausible signaling model, where edges lying on cycles may have unique signs and, hence, may no longer be contracted.

The first variant we consider, A-shortest-path (ASP), considers a signaling model where the length of a satisfying path is always assumed to be the shortest possible. The shortest path assumption is motivated from the observation that signaling pathways tend to be of short length [214]. For each edge $(u, v) \in E$, let $x_{uv} = 1$ denote whether its sign is $-$ (0 if $+$). Similarly, we re-write the signs $\delta_{st} \in \{+, -\}$ as $\delta_{st} \in \{0, 1\}$. Due to the nature of knockout experiments, there are usually much fewer sources compared to targets. Hence, for each source s, we construct a subnetwork $G_s(V_s, E_s)$ such that each edge in this subnetwork lies along

a shortest path from s to one of its targets t. This is done by applying a breadth-first-search starting from each source and target [214]. Furthermore, we denote by $N_s(v)$ the set of neighbors of v in G_s and by d_{sv} the length of the shortest path from s to v. Additionally, for each pair (s, v) in G_s, we define auxiliary variables c_{sv}, r_{sv} where $c_{sv} = 0$ implies that under the selected sign assignment there exists a shortest path from s to v with aggregate sign r_{sv}, i.e., the node pair (s, v) is *satisfied* under the selected assignment. (Note, (s, s) is trivially assumed to be satisfied). We also define E^+, E^- which represent subsets of edges in the ILP with known prior positive and negative signs respectively. Then the following ILP formulation can be used to solve this variant of MSA:

$$
\begin{aligned}
\max \quad & \sum_{st} y_{st} \\
\text{s. t.} \quad & 1 + \textstyle\sum_{u \in \{N_s(v) | d_{sv} = d_{su}+1\}} (c_{su} - 1) \leq c_{sv} && \forall s, v \in V_s \setminus s \\
& r_{sv} = \mathrm{XOR}(r_{su}, x_{uv} | c_{sv} = 0) && \forall s, (u, v) \in \{E_s : d_{sv} = d_{su} + 1\} \\
& c_{st} + y_{st} \leq 1 && \forall (s, t) \\
& r_{ss} = 0, c_{ss} = 0, r_{st} = \delta_{st} && \forall (s, t) \\
& x_{uv} = 0 && \forall (u, v) \in E^+ \\
& x_{uv} = 1 && \forall (u, v) \in E^- \\
& y_{st}, x_{uv}, r_{sv}, c_{sv} \in \{0, 1\} && \forall s, t, u, v
\end{aligned}
$$

The XOR relation between r_{sv}, r_{su} and x_{uv} is conditioned on the value of c_{sv}.

That is, $r_{sv} = r_{su} \oplus x_{uv}$ only if $c_{sv} = 0$.It is linearized as follows:

$$r_{sv} - c_{sv} \leq 2 - x_{uv} - r_{su}$$

$$r_{sv} - c_{sv} \leq x_{uv} + r_{su}$$

$$r_{sv} + c_{sv} \geq x_{uv} - r_{su}$$

$$r_{sv} + c_{sv} \geq r_{su} - x_{uv}$$

Let l denote a layer of G_s such that all nodes belonging to this layer have $d_{sv} = l$. Given a feasible solution to the ILP, if $y_{st} = 1$ we can show that there exists a shortest path from s to t with aggregate sign δ_{st}. Indeed, if $y_{st} = 1$ then $c_{st} = 0$ by the third constraint. This implies that $\sum_{u \in N_s(t) | d_{st} = d_{su}+1} (c_{su} - 1) < 0$. Thus, if t is in layer l of G_s, there must exist a neighbor u of t in layer $l - 1$ such that $c_{su} = 0$. Furthermore, if $c_{st} = 0$, x_{ut} is bound by the XOR constraint to have a sign whose product with r_{su} is δ_{st}. Similarly, if $c_{su} = 0$, there must be a neighbor w in layer $l - 2$ where $c_{sw} = 0$ and $r_{sw} \oplus x_{wu} \oplus x_{ut} = \delta_{st}$. By carefully investigating the constraints applicable to the subsequent layers of G_s (i.e., $l - 3, \ldots, 0$) we find that there must exist a shortest path from s to t such that the product of signs along its edges is δ_{st}. The final two constraints incorporate prior knowledge of signs in the ILP.

The second variant we study, 'A-directed-shortest-path' (AdirSP), additionally assumes each shortest path explaining a pair to be directed from the cause to the effect. It is worth noting that one cannot adapt existing ILP solutions to the orientation and sign assignment problems, as both rely on reducing the input graph

into an acyclic one. This reduction does not work when simultaneously optimizing both. Instead, we simply adapt the ASP formulation above to simultaneously find sign and direction assignments to the network. Specifically, we consider a pair (s, t) to be satisfied by a sign and direction assignment over the network if a directed shortest path from s to t in this assignment has aggregate sign δ_{st}. We call this variant of MSA the 'A-directed-shortest-path' (AdirSP). Let $o_{uv} = 1$ denote whether an edge (u, v) is directed from u to v (0 if from v to u) and let the flow variables f^s_{uv} indicate the existence of a flow from u to v. The flow variables allow computing pair reachability in a directed network. The new ILP is:

$$
\begin{aligned}
\max \quad & \sum_{st} y_{st} & & \\
\text{s. t.} \quad & o_{uv} + o_{vu} = 1 & & \forall (u,v) \in E \\
& f^s_{uv} \leq \sum_{w \in N_s(u) \setminus v} f^s_{wu} & & \forall s, (u,v) \in \{E_s : \\
& & & d_{sv} = d_{su} + 1, d_{su} \geq 1\} \\
& f^s_{uv} \leq o_{uv} & & \forall s, (u,v) \in E_s \\
& a^s_{uv} = (1 - f^s_{uv}) \text{ OR } c_{su} & & \forall s, (u,v) \in \{E_s : d_{sv} = d_{su} + 1\} \\
& 1 + \sum_{u \in N_s(v) | d_{sv} = d_{su}+1} (a^s_{uv} - 1) \leq c_{sv} & & \forall s, v \in V_s \setminus s \\
& r_{sv} = \text{XOR}(r_{su}, x_{uv} | c_{sv} = 0, f^s_{uv} = 1) & & \forall s, (u,v) \in \{E_s : d_{sv} = d_{su} + 1\} \\
& c_{st} + y_{st} \leq 1 & & \forall (s,t) \\
& r_{ss} = 0, c_{ss} = 0, r_{st} = \delta_{st} & & \forall (s,t) \\
& x_{uv} = 0 & & \forall (u,v) \in E^+ \\
& x_{uv} = 1 & & \forall (u,v) \in E^- \\
& y_{st}, x_{uv}, o_{uv}, a^s_{uv}, r_{sv}, c_{sv}, f^s_{uv} \in \{0,1\} & & \forall s,t,u,v
\end{aligned}
$$

The first constraint ensures that each edge has a unique orientation. In some feasible solution, if $f^s_{uv} = 1$, then the second and third constraint ensure that a directed path exists from s to v containing edge (u, v).

Note that the XOR relation that helps determine the sign of an edge now additionally depends on the existence of a flow in that edge. The constraint is

linearized as follows:

$$r_{sv} - c_{sv} - 1 + f^s_{uv} \leq 2 - x_{uv} - r_{su}$$

$$r_{sv} - c_{sv} - 1 + f^s_{uv} \leq x_{uv} + r_{su}$$

$$r_{sv} + c_{sv} + 1 - f^s_{uv} \geq x_{uv} - r_{su}$$

$$r_{sv} + c_{sv} + 1 - f^s_{uv} \geq r_{su} - x_{uv}$$

Another change from the previous formulation is the definition of auxiliary variables a^s_{uv} for each edge participating the ILP. Their value depends on the flow in edge (u, v) originating from s and on c_{su}. The OR relation between these variables is linearized as follows.

$$\begin{aligned} a^s_{uv} &\leq (1 - f^s_{uv}) + c_{su} \\ a^s_{uv} &\geq 1 - f^s_{uv} \\ a^s_{uv} &\geq c_{su} \end{aligned}$$

Given a feasible solution in which $y_{st} = 1$, we show that there exists a shortest path oriented from s to t such that its aggregate sign is δ_{st}. Let t be in layer l of the shortest path graph G_s. If $y_{st} = 1$, then by the seventh constraint $c_{st} = 0$. It follows that $\sum_{u \in N_s(t) | d_{st} = d_{su}+1} (a^s_{ut} - 1) < 0$ (by constraint 5), which implies that there exists a neighbor u in layer $l - 1$ where $a^s_{ut} = 0$. This implies $f^s_{ut} = 1$, $c_{su} = 0$ (constraint 4) and δ_{st} must be the product of the signs given by x_{ut} and r_{su} (constraint 6). Additionally, $c_{su} = 0$ implies there exists a neighbor w in layer $l - 2$ where $a^s_{wu} = 0$

(constraint 5). This implies $f^s_{wu} = 1$, $c_{sw} = 0$ and $r_{sw} \oplus x_{wu} \oplus x_{ut} = \delta_{st}$. In this manner after carefully investigating the constraints through subsequent layers of G_s (i.e $l-3, l-4, ..., 0$) we can find a directed shortest path from s to t such that the product of signs along its edges is δ_{st}. The last two constraints account for signs that are already known.

The underlying assumption in both signaling models above is that a single path is sufficient to force a predefined effect. However, due to the inherent stochasticity in signaling, this might not always be the case [119]. Hence, we strengthen the pair satisfaction assumption in the ASP model to require that a pair (s, t) is satisfied if **all** shortest paths connecting s to t admit the same aggregate sign δ_{st}. We call this variant 'All-shortest-paths' (AllSP) and solve for it using the following formulation:

$$
\begin{aligned}
\max \quad & \sum_{st} y_{st} & & \\
\text{s. t.} \quad & c_{su} \leq c_{sv} & & \forall s, (u,v) \in \{E_s : d_{sv} = d_{su} + 1\} \\
& r_{sv} = \text{XOR}(r_{su}, x_{uv} | c_{sv} = 0) & & \forall s, (u,v) \in \{E_s : d_{sv} = d_{su} + 1\} \\
& c_{st} + y_{st} \leq 1 & & \forall (s,t) \\
& r_{ss} = 0, c_{ss} = 0, r_{st} = \delta_{st} & & \forall (s,t) \\
& x_{uv} = 0 & & \forall (u,v) \in E^+ \\
& x_{uv} = 1 & & \forall (u,v) \in E^- \\
& y_{st}, x_{uv}, r_{sv}, c_{sv} \in \{0,1\} & & \forall s,t,u,v
\end{aligned}
$$

As above, let t belong to layer l of G_s. Given a feasible solution to this new formulation, if $y_{st} = 1, c_{st}$ must be 0 (from third constraint). Hence, for every neighbor u of

t that lies in layer $l-1$ of G_s, $c_{su} = 0$ (from 1st constraint). This in turn constrains the sign assignment of the respective edges (i.e $r_{su} \oplus x_{ut} = \delta_{st}$, for all neighbors u in layer $l-1$). By carefully investigating the constraints through subsequent layers of G_s (i.e $l-2, l-3, ..., 0$), it becomes apparent that for any node v in G_s, all shortest paths from s to v must admit the same aggregate sign (r_{sv}). Hence, all shortest paths from s to t must have an aggregate sign δ_{st}.

Notably, the models discussed above permit mathematically efficient formulations. Specifically, if p is the number of sources ($p << k$), then each formulation contains $O(k + p|V| + |E|)$ variables and $O(k + p(|V| + |E|))$ constraints.

Each of the above models may admit multiple sign assignments with optimal or near optimal scores. Hence, it is necessary to quantify the robustness of a sign assignment to an edge. To this end, we solve each ILP repeatedly n times; each time adding a small Gaussian noise of mean 0 and variance 0.01 to the objective function as shown below. This stochastic approach, motivated by [92], effectively results in a random sampling of different likely solutions that exist nearby in the optimum solution space, thereby allowing us to assess the robustness of the sign on each edge. The procedure is as follows:

1: **procedure** GETSCORES(ILP,n)
2: scores$_{uv} \leftarrow 0$, $\forall (u,v) \in E$ that are in ILP
3: for i = 1:n
4: set objective: $\sum_{st}(1+\epsilon_{st})y_{st}$, where $\epsilon_{st} \sim \mathcal{N}(0,0.01)$
5: $x^* \leftarrow$ solve(ILP)
6: scores$_{uv}$ = scores$_{uv} + x^*_{uv}/n$, $\forall (u,v) \in E$ that are in ILP
7: *return* scores

An edge score close to 1 implies that the sign is negative with high confidence,

a score close to 0 implies a positive sign with high confidence and a score close to 0.5 implies that the sign on that edge cannot be uniquely determined (possibly implicating the absence of an activation/repression effect). For efficiency, we use $n = 10$ throughout. Our conclusions do not change for larger values of n.

2.3 Results

2.3.1 Guilt by association reveals functional heterogeneity in breast cancer

Our overall strategy is to (1) project PIN onto each transcriptomic sample, (2) diffuse functions across the sample-specific PIN to estimate sample-specific function of each gene, and (3) analyze functional changes across conditions. Starting from a previously curated PIN [205], with 16,562 genes and 262,780 edges, we project the PIN on each sample-specific transcriptome, by removing the nodes corresponding to unexpressed or lowly expressed genes (RPKM ¡ 1; see Methods), to obtain a sample-specific PIN. This general approach to obtain a sample-specific network has been used previously to obtain tissue-specific networks in human [19]. For each of the 1184 functional terms (1175 GO terms and 9 NetPath cancer-related pathways, see Methods), in each of the 1157 sample-specific PINs (110 breast cancer samples and 1047 normal breast tissue samples from TCGA [1]), we diffuse the function across the network starting from known annotated (and expressed) genes to yield a raw score for each node. Such sample-specific diffusion-based functional inference across normal and cancer samples allows us to identify specific genes that significantly gain

or lose a particular function in cancer samples, and to assess whether a function has significantly gained or lost genes performing the function in cancer.

After diffusing each of the 1184 functional terms f across 110 normal and 1047 breast cancer samples, we assessed for each gene g whether the fraction of samples in which g is deemed to have the function is significantly different between the normal and tumor tissues based on a Fisher exact test; a greater fraction in cancer is referred to as functional gain and the opposite as functional loss. In addition to statistical significance, we require that the ratio of the fractions of samples where the gene is deemed to have the function in cancer versus normal $\geq \theta$ (gain), or $\leq \frac{1}{\theta}$ (loss). The default value used in the main results is $\theta = 10$ (our conclusions are robust for θ from 2 to 10). We denote by Δf the difference between the number of genes deemed to have gained function f and the number of genes deemed to have lost it. Positive values of Δf indicate net gain and negative values indicate net loss of that function in cancer relative to normal. In total, 732 functions are predicted to undergo a net loss in cancer and 417 are predicted to have a net gain. Table 1 lists the top 10 functions gained and lost. Note that for a function if a majority of genes annotated to have that function are differentially expressed between normal and tumor tissues then Δf will simply reflect this differential expression of the annotated genes and not the effect of altered PIN. To ensure that our inference of functional loss and gain is independent of differential expression of the genes annotated to have the function, when calculating Δf, we exclude the genes annotated with the function. Consistently, as shown in Table 2.1, the functions inferred to have been lost or gained based on Δf exhibit modest log fold changes between normal and cancer in terms

of average number of annotated genes expressed in each cohort, and therefore may go undetected by standard differential expression analysis. Interestingly, overall, we see a weak inverse correlation between Δf and the log fold change based on expressed annotated genes (Spearman correlation = -0.09). Thus our approach uniquely reveals cancer-associated functions. For instance, we find mitotic spindle organization to be lost in cancer consistent with previous reports associating spindle misalignment with cancer [137]. Likewise, we find positive regulation of smooth muscle cell proliferation to be gained in cancer, consistent with prior studies [44].

Table 2.1: Top 10 gained (green) and lost (red) functions are shown, along with Δf, Δf divided (normalized) by the number of genes annotated by the function, and the log fold change, which is the log ratio of the average number of expressed genes annotated by f in cancer and normal samples.

GO ID	Description	Δf	Normalized Δf	log fold change
GO:0048661	positive regulation of smooth muscle cell proliferation	893	15.13	-0.04
GO:0048010	vascular endothelial growth factor receptor signaling pathway	744	10.19	-0.01
GO:0051279	regulation of release of sequestered calcium ion into cytosol	740	13.21	-0.03
GO:1901983	regulation of protein acetylation	723	12.05	-0.04
GO:0000910	cytokinesis	527	6.84	-0.02
GO:0010676	positive regulation of cellular carbohydrate metabolic process	523	8.43	-0.05
GO:0051291	protein hetero oligomerization	508	5.90	-0.03
GO:0042552	myelination	394	6.67	-0.03
GO:2000756	regulation of peptidyl-lysine acetylation	369	6.47	-0.03
GO:0016575	histone deacetylation	333	5.64	-0.01
GO:0006334	nucleosome assembly	-310	-3.13	0.04
GO:0051148	negative regulation of muscle cell differentiation	-127	-2.49	-0.06
GO:0007032	endosome organization	-75	-1.27	-0.007
GO:0018022	peptidyl-lysine methylation	-65	-0.91	0.002
GO:0007052	mitotic spindle organization	-64	-1.054	0.005
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	-56	-0.62	0.003
GO:0016236	macroautophagy	-53	-0.71	-0.01
GO:2000117	negative regulation of cysteine-type endopeptidase activity	-52	-0.61	0.005
GO:0051437	pos reg of ubiquitin-protein ligase activity in regulation of mitotic cell cycle transition	-51	-0.68	0.006
GO:0031145	anaphase-promoting complex-dependent catabolic process	-43	-0.57	0.006

Our analysis above identifies functions with net loss in cancer induced by PIN changes. For such a function f, if a gene g exhibits PIN-induced loss of function

f, then it is likely that mutation-induced loss in activity of g may also be linked to cancer. In other words, for lost functions (negative Δf) we might expect to see more frequent mutations among the genes contributing to the functional loss. We assessed for each function (irrespective of net loss or gain) if it exhibits an elevated mutation frequency among its lost genes (Methods); we explicitly excluded the genes annotated with the specific function. We find that a much greater fraction of lost functions exhibit elevated mutation frequency among their lost genes compared to gained functions used as a control (Fisher p-value = 0.008, odds ratio = 2.36; Table 2; Methods). We repeated this analysis for all values of θ from 2 to 10 and additionally for $\theta = 2$ combined with FDR < 0.1 to ascertain loss/gain of a gene relative to a function. In 9 of the 10 tests, the odds ratio > 1, with an average odds ratio of 1.51. As an alternative, we directly quantified Spearman correlation between Δf and mean mutation rate of corresponding lost genes. Again, in 9 out of 10 cases, consistent with our expectation, we found a weak but significant (all p-values $<$ 0.005) inverse correlation. Likewise, instead of mutations when we use deletion CNV rates to quantify loss in activity (Methods), we find that compared to gained functions, a larger fraction of lost functions exhibited an elevated deletion CNV rate (Table 2.2). While the Fisher test p-value was marginal (0.09), the odds ratio was 2.15. After repeating the tests as above for other values of θ, in 8 of the 10 tests, the odds ratio > 1, with an average odds ratio of 1.59. As an alternative, we directly quantified Spearman correlation between Δf and deletion CNV rate of corresponding lost genes across all functions. In all 10 test cases, consistent with our expectation, we found a weak but significant (all p-values < 0.001) inverse

correlation. These results suggest that a change in network neighborhood of a gene may provide an alternative mechanism for functional loss, in addition to mutations and deletion CNVs.

Table 2.2: The Fisher test contingency table showing the distribution of functions with elevated mutation rates (columns 2 and 3) and deletion CNV rates (columns 4 and 5) between lost and gained functions. Mut(f) = 1 denotes significantly higher mutation rates among the genes contributing to functional loss. CNV(f) = 1 has an analogous interpretation for deletion CNV.

	Mut(f) = 1	Mut(f) = 0	CNV(f) = 1	CNV(f) = 0
Δf <0	48	684	26	706
Δf >0	12	405	7	410

We further assessed whether functions that exhibit cancer-associated gain or loss also exhibit a consistent association with patient survival. For instance, for a function with net loss in cancer relative to normal tissues, we expect that among cancer patients the lower the activity of the function, the worst the patient survival (and the converse for gained functions). To test this association, for each function we estimate its sample-specific activity as the number of genes inferred to be performing that function based on diffusion scaled across all samples. We then estimate the association between patient survival risk and our diffusion-based sample-specific activity of each function using a Cox proportional hazard regression model adjusted for differences in age, and stratified by sex and race. A significant negative (respectively, positive) regression coefficient β corresponds to negative (respectively, positive) association with risk. Of the 1149 functions (732 net loss and 417 net gain), 137 exhibited significant association with survival risk (p-value < 0.05). Of these, 111 were negatively associated with risk, and interestingly, these were significantly biased toward lost functions, consistent with our hypothesis (Table 2.3,

columns 2 and 3; Fisher test p-value = 1.1E-3; odds ratio = 2.1). Only 26 of the 137 were positively associated with risk, but consistently, these were biased toward gained functions (Table 2.3, columns 4 and 5; Fisher test p-value = 5.1E-5; odds ratio = 5.7). As an alternative assessment, we found a significant positive correlation between Δf and β (Spearman correlation = 0.29; p-value < 2.2 E-16). These results suggest that diffusion-based inference of cancer-associated functional change may also be associated with the severity of the tumor among cancer patients. We repeated the above analyses for all values of θ from 2 to 10 and additionally for $\theta = 2$ combined with FDR < 0.1 to ascertain loss/gain of a gene relative to a function. 29 of the 30 tests are consistent with the results above.

Table 2.3: Fisher test contingency table to test for association between functional loss/gain with associations with patient survival; β indicates the association of tumor-specific functional activity with survival risk.

	β<0 & p-value $\leq$ 0.05	p-value >0.05	β>0 & p-value $\leq$ 0.05	p-value >0.05
Δf <0	87	639	6	639
Δf >0	24	373	20	373

Encouraged by the results above, we directly assessed the power of our diffusion-based sample-specific activity profile of a function in predicting patient survival. To this end, we selected the top 1% and bottom 1% (=24) most cancer-associated functions (ordered by Δf), and for each function we estimated its diffusion-based activity in each tumor sample, as defined above. Using the inferred activity levels of these 24 functions as sample-specific features, we then computed the cross-validation accuracy of patient survival prediction based on multivariate Cox regression. The prediction accuracy was quantified using the standard concordance or C-index met-

ric [173]. We find that cross-validation C-index is 0.567. We further included the 9 cancer-related pathways from the NetPath database [109], namely, EGFR1, FSH, IL-1, IL-4, IL-5, Leptin, RANKL, TNF-alpha, and TSH. This extended feature set of 33 functions yielded a cross-validation C-index of 0.62. As a control, we assessed whether the standard alternative approach to quantify sample-specific functional activity, based simply on expressed annotated genes could be equally effective. For candidate features, we assessed the median number of expressed annotated genes in each sample and identified 24 most differentially active functions based on the absolute log ratio of the medians in cancer and normal samples. Adding the 9 NetPath pathways to this list results in 33 features, as above. We then quantified sample-specific activity of these 33 features based on the number expressed annotated genes scaled across all samples and estimated the concordance in an identical fashion to our diffusion-based approach above. This yielded a C-index of 0.51, which is significantly lower than 0.62 (p-value = 0.01). We repeated the above analyses for all values of θ from 2 to 10 and $\theta = 2$; FDR < 0.1 to ascertain loss/gain of a gene relative to a function. In all cases the diffusion-based C-index is higher than the control, and significantly so in 8 out of 10 cases. We further validated the survival prediction accuracy of our diffusion-based functional activity profile in an independent METABRIC breast cancer dataset [53]. We used the features derived from TCGA dataset as above and used those to assess cross-validation prediction accuracy of the diffusion-based and annotation-based methods in METABRIC. Again, we find that C-index of the diffusion-based approach was 0.62 whereas the annotation-based approach achieved an accuracy of 0.57 (difference p-value = 0.0004). These consis-

tent results across datasets suggest that the diffusion-based approach to quantify functional activity may provide additional information about the functional state of a tumor, relevant to patient survival.

We further tested if our novel diffusion-based functional activity profile is predictive of known clinical characteristics of breast tumors, specifcially, the cancer subtype (Basal, Her2, Luminal A, Luminal B, Normal), and its hormone response status, Estrogen Receptor positive (ER+) and Progesterone Receptor positive (PR+). Based on clinical annotation of the METABRIC tumors, we trained 7 different Support Vector Machine (SVM) models, one per clinical indicator, using randomly selected 50% of the samples to train and the other half to assess the prediction accuracy, quantified by ROC-AUC. We repeated the training and testing 2000 times to obtain mean and 95% confidence interval. Note that while the training and testing of the model is done on METABRIC, the cancer-associated functions used as features were inferred from TCGA data independently. We compared the performance of our diffusion-based functional activity profile with annotation-based activity profiles as above. Table 2.4 shows the AUC estimates of each model. We found in all classification tasks, the diffusion-based model can predict each clinical indicator more accurately than the alternative annotation-based approach.

Table 2.4: The following table displays the AUC estimates of the 7 independent classifiers trained with two different feature sets (diffusion-based functional activity and annotation-based functional activity) for each clinical indicator

Clinical Indicator	AUC - Diffusion	AUC – Annotation (Control)
Basal	0.91 (95% CI = 0.887-0.928)	0.88 (95% CI = 0.842-0.893)
Her2	0.77 (95% CI = 0.737-0.808)	0.72 (95% CI = 0.672-0.747)
Luminal A	0.79 (95% CI = 0.763-0.806)	0.76 (95% CI = 0.745-0.788)
Luminal B	0.78 (95% CI = 0.752-0.8)	0.75 (95% CI = 0.735-0.786)
Normal	0.72 (95% CI = 0.685-0.761)	0.69 (95% CI = 0.64 -0.724)
ER+	0.93 (95% CI = 0.916 – 0.949)	0.87 (95% CI = 0.857-0.899)
PR+	0.77 (95% CI = 0.742-0.784)	0.75 (95% CI = 0.731-0.774)

Next, using our diffusion-based activity profiles of the 33 functions (24 GO terms and 9 cancer-related NetPath pathways) used above, we clustered all METABRIC samples using Nonnegative Matrix Factorization (NMF) [25], in an unsupervised fashion, into 10 groups (Methods). Figure 2.3 panel A shows that the distribution of the five known subtypes of breast cancer across the 10 clusters. Even though the functional profile-based clustering are not associated with known subtypes, interestingly, as seen in Figure 2.3 panel B, the diffusion-based unsupervised clusters exhibit significant inter-cluster differences in patient survival (Log rank p-value = 3.2E-3). In contrast, when we use annotation-based functional activity profiles to cluster the tumors following an identical procedure as above, the clusters did not reveal a difference in survival across clusters (Log-rank p-value = 0.23). We fitted a Cox proportional hazards model to the METABRIC survival data using cluster membership as a feature while controlling for age, sex and race, as above. Cluster memberships generated by diffusion-based functional activity profiles show a significant association with survival risk (β = 0.04, p-value = 8.5E-3) whereas cluster memberships generated by annotation-based profiles had no significant effect (β =

0.01, p-value = 0.32). These results suggest that in addition to expression based changes, PIN-induced functional changes of genes in breast tumors may also play a functional role in cancer.

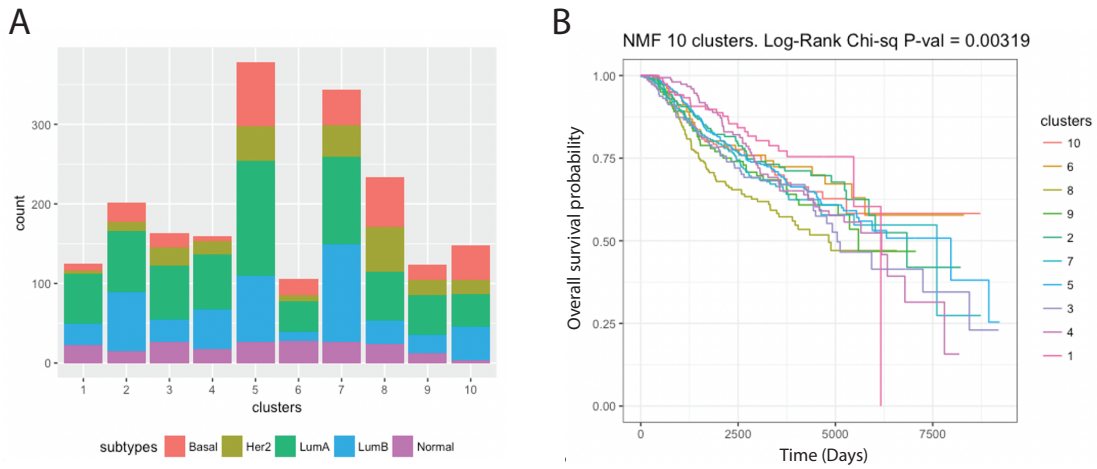

Figure 2.3: (A) Distribution of five known subtypes across 10 clusters inferred from diffusion-based activity profiles. (B) Kaplan-Meier survival curves of patients grouped in the 10 clusters show significant survival differences.

Figure 2.4 shows, for each of 7 subtypes, the log-fold change in average functional activity of the 33 functions (24 GO processes and 9 Netpath pathways) in samples corresponding to the subtype versus the rest. The most notable changes are increase in activity of ovulation cycle process (GO:0022602), Epidermal Growth Factor Receptor signalling pathway (EGFR1), and Receptor Activator of Nuclear factor Kappa-B Ligand signalling pathway (RANKL) in ER+ breast tumors. Previous experimental and clinical studies have shown that increased level of EGFR in ER+ breast tumors leads to resistance to hormone therapy [153, 70] through hormone independent activation of estrogen receptors [28]. As seen in Figure 2.4, the EGFR1 signalling pathway has a 0.23 log fold higher average functional activity in ER+ breast cancer patients (70% of which were recorded to have taken hormone

therapy). This suggests that PIN-induced increase in EGFR signalling activity in ER+ tumors may lead to increased levels of EGFR thereby increasing the possibility of hormone therapy resistance. Our results also indicate a 0.24 log fold higher functional activity in RANKL signaling in ER+ breast cancer. While RANK and RANKL are normally expressed in mammary gland epithelial cells, they are also expressed in many epithelial breast tumor cells. RANKL has been experimentally shown to induce cell migration in epithelial tumor cells expressing RANK, and is also an important osteoclast differentiation factor found highly expressed in the bone marrow thereby creating a conducive environment for bone specific metastasis [107]. This is consistent with the observation that many tumors in breast that are known to recur in bone tissue are ER+ [101]. Moreover, inhibition of RANKL expression in combination with hormone therapy has been shown to improve treatment efficacy and prevention of bone metastasis in experimental mouse models of ER+ tumors [37]. These results suggest that the knowledge of PIN guided functional changes in genes via guilt by association may provide important biological insights into mechanisms of treatment resistance.

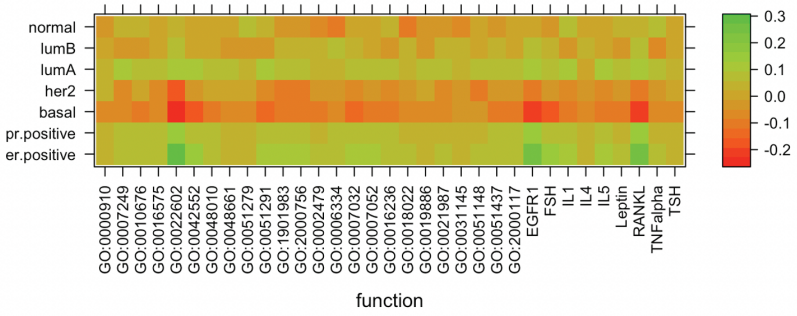

Figure 2.4: The following figure displays the log ratio between the average numbers of genes assigned to each function by diffusion (represented by columns) across samples annotated with a particular subtype (represented by rows) versus the rest of the samples.

2.3.2 Imputing functional effects of protein-protein and protein DNA interactions in Yeast

We focused our analysis on budding yeast (Saccharomyces cerevisiae). We obtained 4095 protein-DNA interactions spanning 2079 proteins (conserved across at least 2 other yeast species) from [136]. We additionally downloaded 2930 high-quality experimentally verified protein-protein interactions from [254], 1361 kinase-substrate/phosphatase-substrate interactions (KPIs) among 802 proteins from [31], and 189 physical interactions from signaling pathways of yeast in KEGG. We merged these sets into a *unified* yeast network of 8268 unique physical interactions among 3695 proteins.

We extracted all 110,487 knockout pairs spanning 6228 proteins from [182] and additionally 699,771 pairs spanning 6110 proteins from [113]. A pair was assigned a positive sign if the target gene was repressed in response to knockout of the source, and a negative sign if the target gene was activated/up-regulated. We restricted ourselves to knockout pairs such that the absolute log fold change in expression of the target gene is > 2 and FDR < 0.001. This leaves us with 1756 significant knockout pairs from [182], referred to here as the *Reimand set*, and 3524 significant knockout pairs from [113], referred to here as the *Kemmeren set*. The above choice of thresholds was made while taking into consideration the inherent computational complexity of the problem.

For a systematic validation of our sign prediction models we collected sign information as follows. 147 of 192 physical interactions in yeast had an exper-

imentally confirmed sign from KEGG (See Figure 1).In addition, following[97], we extracted GO molecular function annotations related to transcriptional activators (GO:0045893) and transcriptional repressors (GO:0045892). Protein-DNA interactions originating from transcriptional activators were given a positive sign whereas protein-DNA interactions originating from transcriptional repressors were given a negative sign. Finally, we also extracted information on protein kinases (GO:0004672) and protein phosphatases (GO:0004721). We reasoned that since there are roughly 3 times as many confirmed functionally activating phophorylation sites compared to repressive ones (PhosphoNET database, `www.phosphonet.ca`), and that 71% of phosphorylation interactions of yeast in KEGG are annotated as activating and 81% of de-phosphorylation interactions of yeast are annotated as repressing, kinase-substrate interactions tend to be activating while phosphatase-substrate interactions tend to be repressing. Thus, physical interactions linking a GO annotated kinase and a substrate were given a positive sign whereas interactions linking a GO annotated phosphatase to a substrate were given a negative sign. Any interaction in the unified network that had conflicting signs was left unsigned (unless it had sign information from KEGG, in which case this latter information was used). In summary, the validation set consists of three groups of signed interactions in the network: (i) 2014 (1131 +, 883 −) signed protein-DNA interactions, (ii) 1044 (872 +, 172 −) signed kinase/phosphatase-substrate interactions, and (iii) 147 (96 +, 51 −) signed KEGG interactions.

We evaluated each of the four models presented above in a 5-fold cross-validation setting on the unified yeast network, focusing on the interactions covered

by each model, i.e., participating in the corresponding ILP. To this end, we randomly divided all signed and covered interactions into 5 equal parts.Using each model, we predicted the activation/repression potential of the interactions in each part while constraining the signs of interactions in the remaining parts. Then we measured the performance of the activation/repression scores of a given model across the five parts for different subsets of signed interactions covered by the model. For each subset, we denote its set of covered positive and negative interactions by E^+ and E^-, respectively.

As a benchmark, we discuss the performance of the previous A-path model. Recall that in this model we should contract all interactions that lie in a strongly signed block of size ≥ 3. Since all blocks were strongly signed, this resulted in an acyclic network with 77% of the interactions contracted. When working with knockout pairs from the Reimand set, we observe that only 1% of all the network interactions participate in the ILP constraints due to network reduction, and 25 of them belong to the validation set. Due to low coverage over the validation set, we instead evaluated this framework using knockout pairs from the Kemmeren set. Overall, 4% of network interactions are covered in this instance and 73 interactions from the validation set were part of the ILP formulation, yielding an AUC of 0.66. Since there were only 73 interactions to validate our predictions, we could not evaluate the performance on individual subsets.

Next, we evaluated the ASP, AdirSP and AllSP models over the unified network. Tables 2.5 and 2.6 summarize the performance over the validation PDIs, KPIs, and the KEGG interactions. We find that our new formulations lead to sign

assignments on 35% of network interactions when working with the Reimand set and 59% of network interactions when working with the Kemmeren set; $\approx$ 15-fold coverage increase compared to previous work (A-path).

Table 2.5: Performance evaluation of different imputation models on the Reimand set (coverage of 35%).

interaction	$\lvert E^+\rvert, \lvert E^-\rvert$	AUC (ASP)	AUC (AdirSP)	AUC (AllSP)
PDI	435, 458	0.75	0.63	0.84
KPI	205, 20	0.83	0.56	0.72
KEGG	40, 27	0.56	0.52	0.65

Table 2.6: Performance evaluation of different imputation models on the Kemmeren set (coverage of 59%).

interaction	$\lvert E^+\rvert, \lvert E^-\rvert$	AUC (ASP)	AUC (AdirSP)	AUC (AllSP)
PDI	744, 653	0.63	0.59	0.83
KPI	522, 98	0.61	0.51	0.77
KEGG	46, 32	0.58	0.54	0.71

In order to directly compare the performance of the A-path model to our suggested alternative models, we evaluated them on the restricted validation set of 73 interactions covered by the A-path model.On this set ($|E^+| = 49, |E^-| = 24$) the performance of AdirSP was lower to A-path (AUC of 0.64), while ASP and AllSP had better performance (AUCs of 0.73 and 0.68, respectively)

Previous work as well as our models above vary in the assumptions they make on the way a knockout effect is explained, going all they way from requiring a single path of any length to requiring all paths of shortest length. Note that we adopt these models partly because they are grounded in our very own observations of cellular

signaling pathways and because they permit an efficient mathematical formulation. These descriptions are not perfect. In turn, the solution of each model allows different degrees of freedom on the signs of underlying interactions. To make the best inference possible for each physical interaction given the complex nature of cellular signaling, we integrate the predictions of each model in an ensemble. That is, using the sign scores from solutions to ASP, AdirSP and AllSP as features, we train a hybrid model, specifically a random forest classifier, that makes an overall prediction of the sign of an interaction (A-path was excluded due to low coverage). The ensemble model is evaluated via nested cross-validation. In detail, the validation set is divided into the same 5 parts as above. Four of the parts are used for training the individual models to score the fifth part. Next, we perform a 5-fold cross validation on the fifth part to train and test the classifier. Finally, using the cross-validated predictions across all parts, we report the mean classifier performance (AUC) against the signs of different validation subsets. Tables 2.7 and 2.8 summarize the performance of the random forest classifier on the different knockout sets and validation subsets.

Table 2.7: Performance evaluation of the random forest classifier using the Reimand set.

interaction	$\|E^+\|, \|E^-\|$	AUC (classifier)
PDI	435, 458	0.86
KPI	205, 20	0.85
KEGG	40, 27	0.77

Table 2.8: Performance evaluation of the random forest classifier using the Kemmeren set.

interaction	$\|E^+\|, \|E^-\|$	AUC (classifier)
PDI	744, 653	0.80
KPI	522, 98	0.67
KEGG	46, 32	0.81

Overall, we observe that the classifier outperforms all individual models on the set of curated interactions from KEGG.It also outperforms the different models with respect to PDIs and KPIs on the Reimand set. The lower performance of the classifier on the KPI set (compared with the AllSP model) when working with the Kemmeren set is likely an artefact resulting from the skewed distribution of class labels. Such a skew may influence ensemble classifier performance on unseen data.

2.4 Discussion

This work demonstrates that tumors can exhibit transcriptional heterogeneity to adapt and survive in harsh environments. This transcriptional heterogeneity can potentially lead to gain or loss of function of genes via "guilt-by-association" that provide tumors with a fitness advantage. Using our diffusion based approach one can uncover such events. However, this approach has a few notable limitations. First, the guilt-by-association is a trend and there are several exceptions to the general principle, as described previously [76], and second, the diffusion algorithm is effective for relatively large functional groups. We have explicitly addressed these limitations by restricting our analysis to those functional groups that yield a rea-

sonable diffusion-based recall, suggesting that these functions are broadly clustered in the PIN, and by only considering functional groups with at least 50 genes (and at most 500 genes, as discussed in Methods). It is interesting to note that the number of genes implicated in a function can far exceed the number of genes currently annotated by the function, consistent with substantive incompleteness of functional annotations. However, it is difficult to verify these predicted functional implications, except indirectly through their predictive value in various tasks, as we have done or via high throughput genetic interaction screens. Follow up investigations will provide further insights and strengthen our conclusions. For instance, it will be instructive to first experimentally test in model organisms such as yeast, the extent to which context-specificity of genetic interactions can be explained by functional re-wiring of the PIN.

In addition to the above, we developed novel mixed integer linear programming models to infer the flow of biological signals over protein-protein interaction networks. We discussed the underlying assumptions guiding the predictions of each model and its advantages in terms of coverage relative to prior work by [97]. We then measured the cross-validation accuracy of each in predicting signs across two knockout datasets in yeast to find that our models lead to improvement in accuracy and coverage over the previous state-of-the art method by [97]. We eventually train a hybrid signaling model based classifier that learns to best combine predictions of each model. This was partly motivated by the fact that the three models presented in this work, although mathematically efficient to represent, are insufficient to capture the true complex nature of cell signaling.Furthermore, this warrants the

exploration of other plausible models that could be potentially integrated into the classifier to improve its predictions. The github code related to this work is available at: `https://github.com/spatkar94/NetworkAnnotation.git`.

Chapter 3

Intra-tumor genetic heterogeneity (ITH) and its impact on response to immune checkpoint blockade therapy

⋆⋆ This work was done in collaboration with Dr. Yardena Samuels and now appears in Cell [249]

3.1 Overview

It has recently been shown that immunotherapy strategies that enhance anti-tumor T-cell response, such as checkpoint inhibitors and adoptive T-cell therapy, exhibit remarkable clinical effects in a wide range of tumor types [186, 248]. However, many tumors do not respond to checkpoint inhibitors and the determinants of treatment efficacy remain largely unknown [208]. Neo-antigens that arise as a consequence of somatic mutations within the tumor represent an attractive means to promote immune recognition in cancer [87]. Indeed, high TMB and neo-antigen load in tumors have been associated with an enhanced response to immune checkpoint blockade therapy [73, 95, 204, 217, 233]. Cutaneous melanoma, which is among the most highly mutated malignancies [4], has the highest objective response rates to checkpoint blockade (60% upon combined CTLA4 and PD-1 blockade) [121]. There is a growing appreciation of the key role of T-cell mediated responses against neo-antigens in mediating responses to melanoma therapy [50, 83, 87, 222], as well as

the characterization of T cell activation and dysfunctional states [45, 118, 127, 201].

While the leading hypothesis in the immunotherapy field is that tumors with increased TMB present more neo-antigens and, thus, are more immunogenic [73, 85, 95, 192, 222, 233], tumors containing equally high TMB exhibit a variable immune response [199], and some cancers with low TMB can respond to immunotherapy [148], thus again questioning the association between TMB and response. Moreover, predicted neoantigen load does not correlate with T cell infiltration in melanoma [218], and TMB alone is not a sensitive or specific predictor of outcome to treatment [99], suggesting that additional factors determine the development of T-cell reactivity.

In parallel, it has recently been reported that ITH, manifested by the distribution of clonal vs. sub-clonal mutations and neoantigens [144, 218], may influence immune surveillance [142, 143, 184] and pan cancer analyses show better survival for tumors with low ITH [7, 149, 150, 156]. Despite past attempts to model the effect of increased TMB [239] or ITH [71], no attempts were made to study effects of TMB and ITH on immune response in a comparative, causal manner. Here we evaluate the contributions of different aspects of ITH and TMB in immune-mediated tumor rejection in mouse models and study its parallels in patient data.

3.2 Methods

3.2.1 Inference of ITH in melanomas from TCGA

From the TCGA data access portal (https://portal.gdc.cancer.gov/), we downloaded level 2 SNP array and germline + somatic variant call data for 432 skin cutaneous melanoma tumor and matched normal samples. Across all 432 patients, we apply CHAT [126] under default package settings to estimate tumor purity followed by estimation of cellular abundance of CNVs and somatic mutations from the SNP array and variant call data respectively. We found that the average sample purity estimated by CHAT is 74% with only 14 samples having purity less than 25%. However, we do not pre-filter any of these samples in our downstream survival analyses as our final conclusions remain the same even after their removal. To estimate mutation burden (TMB) per sample, we count the number of somatic variant calls that were classified as missense or non-sense per sample. This data was obtained from the cbioportal website (https://www.cbioportal.org). Since CHAT detects CNVs using the circular binary segmentation algorithm [167], which essentially partitions the genome into non-overlapping sections of same copy number, we estimate CNV load per sample in a manner similar to [7]. To elaborate: for a tumor sample, let L_s be the length of a segment s of the genome and let CN_s be total copy number of that segment inferred by CHAT while considering the tumor purity. Let $X_s \in \{0, 1\}$ be an indicator of deviation of CN_s from normal diploid copy number of 2. Then,

CNV load is defined as:

$$\text{CNV Load} = \frac{\sum_{s:X_s=1} L_s}{\sum_s L_s} \tag{3.1}$$

Given that tumor evolution is characterized by a series of clonal expansion events, we often find that mutations and CNVs detected from a bulk tumor sample group into clusters. The number of these clusters or clones is interpreted as the intra-tumor heterogeneity. Using CHAT, we can derive two estimates of number of clones by – clustering cellular abundances of somatic mutations (ITH1) or clustering cellular abundances of CNVs (ITH2). Both estimates convey important information of the underlying clonal structure at different resolutions. Hence, we set the overall intra-tumor heterogeneity of a sample as:

$$\text{ITH} = \max(\text{ITH1}, \text{ITH2}) \tag{3.2}$$

Given the limitation of a single bulk tumor sample per patient for inference, the above estimate is a lower bound and correlated with tumor purity (spearman's rho = 0.232, p value = 2.09E-5). However, we show that our downstream results still hold after correcting for tumor purity, age and stage (see below).

To see if mutation burden, CNV load and ITH are associated with overall patient survival, we stratified the TCGA patients into the following groups:

- Low mutation burden ($\leq$median), high mutation burden ($>$ median)
- Low cnv load ($\leq$median), high cnv load ($>$ median)
- Low ITH ($\leq$median), High ITH ($>$ median)

Given a cohort of 402 patients with available clinical data, we then fit Kaplan-Meier survival curves for each group and their combinations and test if there are any significant survival differences between the groups using a Log-rank test. All survival analyses were performed using the survival package readily available for R. Due to potentially confounding effects of purity and other clinical factors, it is necessary to ascertain whether the observed associations with survival still hold after accounting for confounding factors. The three major potentially confounding factors are tumor purity, patient age and clinical stage. We hence performed a multivariate cox regression analysis in which patient age, tumor purity and clinical stage were included as additional factors. Our original conclusions do not change after running this analysis.

3.2.2 Quantifying host immune response from bulk RNA-seq data

Single-end RNA Seq data from the mouse cell-line derived tumors was trimmed using Trimomatic (0.36) to filter out low quality and adaptor reads. The trimmed data was then processed using Salmon (0.9.2) to directly quantify gene expression levels (TPM). Furthermore, gene expression levels (RPKM) of the 432 melanoma patients with corresponding survival information were downloaded from the TCGA portal. Cytolytic activity (CYT) of TILs in the mouse cell-line derived tumors, and likewise in patient tumors, was estimated from the geometric mean of expression levels of GZMA and PRF1 [199].

$$\text{CYT} = \exp(\frac{\log(\text{GZMA} + 1) + \log(\text{PRF1} + 1)}{2}) \tag{3.3}$$

3.2.3 Phylogenetic analysis of mouse UVB and Sngle Cell Clones

Exome sequencing data for the UVB exposed sample (n = 1) and the individual single cell clones (n = 20), were used for joint clustering to infer the subclones present across this combined set of samples. MAF files (generated as described above) and somatic copy number alteration logR scores by segment (generated using CNVkit), were utilized as input to the SciClone clustering algorithm [151]. To ensure high confidence clonal markers were used, the following variant filters were applied: i) a minimum alternative read depth of > 5 was used, ii) indels and triallelic sites were excluded, and iii) only variants present in $\geq$ 2 samples were retained (i.e., private mutations only in one sample were excluded). This latter criteria of filtering out private variants was implemented to minimize the impact of technical artifacts, which are known to be a potential issue in ITH analyses [212], as well as the fact that variants found only in one sample offered minimal utility in inferring the overall cross-sample phylogeny. For completeness, the proportion of private variants found in the experimental mixes used in Figure 6 is included in Table S7, and further studies with high depth error corrected sequencing will be required to accurately understand the biological role of private mutations. SciClone was run with the following parameters: copyNumberMargins = 0.5, maximumClusters = 30 and minimumDepth = -1 (variants were already pre-filtered for minimum

depth of > 5 alternative reads during MAF file creation). The clustering solution from SciClone was manually reviewed, and any obviously poor quality clusters were removed (e.g., clusters defined by < 10 mutations, clusters present in every sample but with low VAF values ($< 25\%$), duplicated clusters). Phylogenetic trees, and representative sample tumor diagrams were constructed using R package CloneEvol [54]. Individual single cell clones were mapped to terminal clones/branches (from the overall clustering solution), based on the closest fitting VAF frequency.

3.2.4 Analysis of human immune checkpoint blockade datasets

Four malignant melanoma cohorts were analyzed, from previously published studies by Snyder et al. (anti-CTLA4 treated), Riaz et al. (anti-PD1 treated), Hugo et al. (anti-PD1 treated) and Van Allen et al. (anti-CTLA4 treated). Pyclone clustering results for the Riaz et al. cohort were obtained directly from the authors supplemental data files (https://github.com/riazn/bms038_analysis/tree/master/data), and clones defined by $n \geq 2$ mutation were retained for further analysis. Pyclone clustering results for the Snyder et al. (2014) [217] and Van Allen et al. (2015) [233] cohorts were obtained from previously published work from McGranahan et al. (2016) [143], with clones already having undergone quality control filtering. For Hugo et al. (2016) [99], no previously published clustering results were available, and instead we managed to successfully process raw WES data of a subset of 22 samples for which there is available survival information on 21 samples. The processing pipeline used is as follows: we called variants for each cancer and paired normal sam-

ples using the GATK (V. 3.6) 'HaplotypeCaller' [128, 145](Li et al., 2009, McKenna et al., 2010) utility applying '-ERC GVCF' mode to produce a comprehensive record of genotype likelihoods for every position in the genome regardless of whether a variant was detected at that site or not. The goal of using the GVCF mode was to capture confidence score for every site represented in a paired normal and cancer cohort for calling somatic mutation in cancer. Next, we combined the paired GVCFs from each paired cohorts using GATK's 'GenotypeGVCFs' utility yielding genotype likelihood scores for every variant in cancer and the paired normal sample. Next, we used GATK's 'VariantRecalibrator' utility using dbSNP VCF (v146: `ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b146_GRCh38p2/VCF`) file by selecting annotation criteria of QD;MQ;MQRankSum;ReadPosRankSum;FS;SOR, followed by GATK's 'ApplyRecalibration' utility with 'SNP' mode. Next, using GATK's 'VariantFiltration' utility we selected the variants with VQSLOD ≥ 4.0. Finally, somatic mutations were defined as the loci whose genotype ($1/1, 0/1,$ or $0/0$) with 'PL' (Phred-scaled likelihood of the genotype) score = 0, i.e., highest confidence) in cancer is distinct from that in paired normal. The final somatic mutations were mapped on an exonic site of a transcript by 'bcftools' tool (V. 1.3)[128] using BED file of coding region. Clustering analysis was then completed using the CHAT algorithm, as described above.

For each case, the count of mutations within each cluster (clone) as defined by Pyclone/CHAT were computed, and Shannon diversity index (SDI) was calculated using the entropy.empirical function in R package 'Entropy'. Overall survival data was obtained from the original author's publications, and n = 3 cases from the Riaz

et al. cohort with "NE" RECIST coding were excluded, on account of death having occurred prior to disease assessment. In addition, the group of n = 10 cases from the Van Allen et al. (2015) cohort with long-term survival but no clinical benefit from anti-CTLA4 treatment were excluded, as per the original publication. All other cases with available survival data and clustering results were used for survival analysis (cases with available clustering results were those with data deposited in `https://github.com/riazn/bms038_analysis/tree/master/data` (Riaz et al., 2017)[185] and `https://bitbucket.org/nmcgranahan/clonalneoantigenanaly-\sispipeline/downloads/` (Van Allen et al., 2015 and Snyder et al., 2014 cohorts) [217, 233], extracted on date 14/05/2019, please refer to the original publications for further details). Kaplan-Meier plots were drawn using the ggsurvplot function in R, with the low/high diversity groups being defined by having a SDI value $<$ or $\geq$ to the median value in each cohort respectively. Significance values in Figure 7 were calculated using the coxph function in R, with SDI included in the model as a continuous variable, and overall survival hazard ratios are reported per unit increase in SDI score. To correct for purity, a multi-variable coxph model was used, with SDI and purity included as variables, and the significance values of each variable in the model were analyzed. Meta-analysis of significance across the two studies was calculated using the Fisher method for combinig p values.

Statistical analyses on immune checkpoint blockade treated datasets were performed using the Prism 8 software (GraphPad, San Diego, CA, USA) and the software environment R, using RStudio. For all statistical analysis a p value of < 0.05 was determined to be significant. All data is presented using standard error mean (SEM).

P values are depicted in all figures, and selected p values with exceptional significance to the paper are also briefly described in the main text. Samples sizes (n), means and SEM are depicted in the figures and/or figure legends. Sample size values were either depiction of number of mice used for experiments, or number of patients. For the comparison of patient survival curves (Kaplan-Meier curves) the log rank test was used. For samples with distribution other than normal, or with small sample size (n ¡ 6), the nonparametric Wilcoxon test, Mann-Whitney's U test, and Kruskal-Wallis test were used. For samples which approximate normal distribution, Student's t test or one-way ANOVA followed by Bonferoni's post hoc test was used. For correlation between CYT score and the number of clones depicted in Figure 3.1, the Spearman's Rho nonparametric test was used. For tumor growth curve, repeated-measures two-way ANOVA was used, followed by Bonferoni's post hoc test. For the analysis of the Shannon diversity index (SDI), z-test from Cox proportional hazard mode was used with SDI tested as a continuous variable. Proportions of genomic mutation types of the different cell lines were analyzed using the Chi-Square test.

3.3 Results

3.3.1 Impact of ITH on patient survival in melanoma

We analyzed a cohort of 402 pre-treatment samples of TCGA [1] melanoma patients with matched genomic and survival outcome information. Patients were grouped based on their mutation burden, copy number variation (CNV), and ITH (estimated as the number of clones), which were computed based on each sample's

somatic copy number alterations and somatic mutation data (See section 3.1). Neither mutation burden nor CNV load, as a single component, was significantly associated with patient survival (Figure 3.1, panels A and B). However, patients with low ITH had significantly better survival (Figure 3.1, panel C), consistent with previous observations (Brown et al., 2014, Morris et al., 2016). Indeed, when patients were segregated by number of clones, distinct survival curves could be seen; patients with low ITH levels (2 clones) had the best survival rate, whereas those with high ITH levels (6 clones) had the worst survival rate (Figure 3.1, panel D). When combining all three factors, we found that patients with a high ITH and a low mutational or CNV load had the worst survival rate (Figure 3.1, panels E and F). These conclusions hold when controlling for potential confounding factors, including age, tumor stage, and tumor purity (See Table 3.1). Finally, for each patient we computed the "cytolytic score (CYT)" [199], which is associated with the degree of anticancer immunity based on the geometric mean expression of two key cytolytic effectors, Granzyme A and Perforin1, which are upregulated upon CD8+ T cell activation and upon effective immunotherapy treatment. CYT scores were significantly higher in patients with low ITH compared with those with high ITH (Figure 3.1 panel G; Wilcoxon rank-sum test, $p = 4.32 \times 10^6$). Notably, the CYT scores were inversely correlated with the degree of number of clones throughout the TCGA cohort (Figure 3.1 panel H; Spearman's rho $= -0.27$, $p = 4.3 \times 10^6$). Together, our results suggest that ITH plays an important role in shaping melanoma host immune response and patient survival.

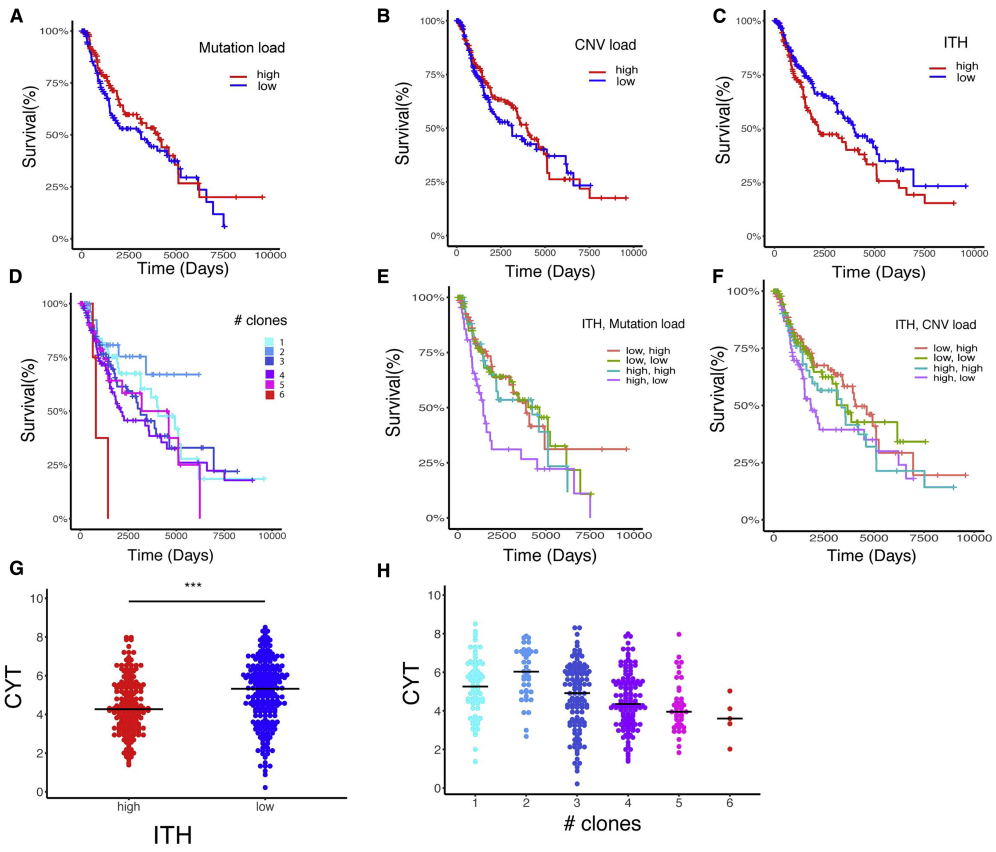

Figure 3.1: (A) Kaplan Meier survival curves (time is measured in days on the x axis) of patients with high versus low mutation burden. Log rank statistics: 1.96, p = 0.16. (B) Kaplan-Meier survival curves of patients with high versus low CNV load. Log rank statistics: 0.31, p = 0.577. (C) Kaplan-Meier survival curves of patient with high versus low ITH. Log rank statistics: 3.97, p = 0.046. (D) Kaplan-Meier survival curves for patients segregated by their number of clones. (E) Kaplan Meier survival curves of patients segregated based on the combination of mutation burden and ITH. Log rank statistics: 9.2, p = 0.0267. (F) Kaplan-Meier survival curves of patients segregated based on the combination of CNV load and ITH. Log rank statistics: 4.57, p = 0.206. (G) CYT score (in log scale) of patients with high versus low ITH. $\star\star\star$p < 0.001, Wilcoxon's test.(H) CYT score (in log scale) of patients segregated by their number of clones. Spearman's rho: -0.27, p < 0.001.

3.3.2 Tumors with lower ITH are swiftly rejected by immuno-competent mice independent of tumor mutation burden levels

$\star\star$ Experimental analyses done by the Yardena Samuels' lab

Following these retrospective association results in human patients, we sought

to establish an experimental in vivo mouse system that would enable us to uncouple TMB and ITH and study their influence on tumor immunogenicity in a causal, systematic manner. First, to assess the effect of increased mutational load and increased concomitant heterogeneity on anti-tumor immunity, we exposed the mouse melanoma B2905 cell line [170] to UVB irradiation (Figure 3.2, panel A), a key carcinogenic source driving melanoma initiation [57]. Because the literature regarding UVB research in melanoma varies considerably with respect to the amount of radiation exposure needed to induce melanoma genesis, we first titrated the amount of radiation needed for an optimal UVB response without compromising cell longevity. We found that a UVB dose of 600 J/m2 was sufficient to induce p53 elevation [34] and cyclobutane pyrimidine dimer (CPD) formation [40] while maintaining the longevity of the murine melanoma cell lines B2905 and B16F10.9

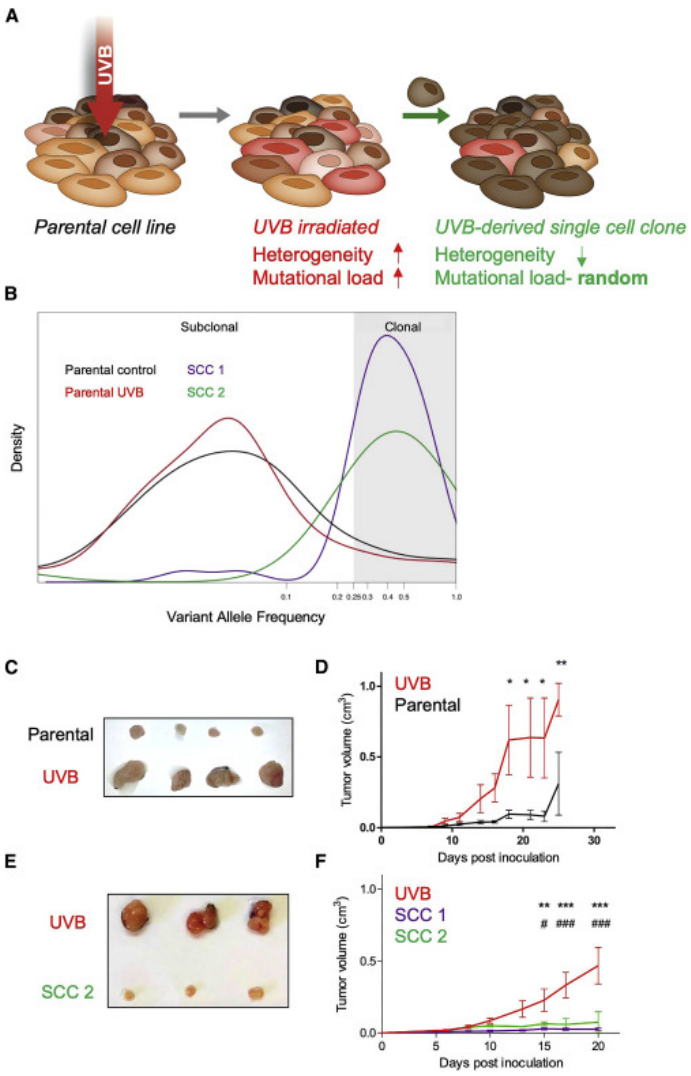

Figure 3.2: (A) Scheme of experimental design for generating UVB-irradiated cells and generating SCCs derived from UVB-irradiated cells. Cell lines are irradiated by UVB at dosage of 600 J/m2; from these irradiated cells, SCCs are generated. (B) Distribution of variant allele frequencies (VAFs) of parental B2905 cells (black), UVB-irradiated B2905 cells (red), SCC 1 (purple), and SCC 2 (green) in log2 space. VAF $> 0.25(\log_2 = 2)$ is considered clonal. (C) Tumors excised from mice inoculated with either parental or UVB-irradiated cell lines on day 15 after inoculation.(D) In vivo tumor growth in mice inoculated with parental B2905 cells (black) and UVB-irradiated cells (red). n = 3–4; data are representative of three independent experiments. Data are mean ±SE. $\star p < 0.05, \star\star\star p < 0.001$, two-way ANOVA followed by Bonferroni's post hoc test.(E) Tumors excised from UVB-irradiated B2905 cells versus SCC 2, day 19.(F) In vivo growth of tumors in mice inoculated with UVB (red) or SCC 1 (purple) and SCC 2 (green). n = 4–5; data are representative of two independent experiments. Data are mean ±SE. $\star\star p < 0.01, \star\star\star p < 0.001$, two-way ANOVA followed by Bonferroni's post hoc test. $\star$ refers to UVB and SSC 1 comparisons; # refers to UVB and SSC 2 comparisons.

In parallel with the increase in TMB upon UVB irradiation, we also detected an increase in ITH from the distribution of the variant allele frequency (VAF; the fre-

quency of a mutation within the population plotted against the probability density function), which was skewed toward a more subclonal phenotype (VAF < 0.25) [245] and exhibited a relatively small fraction of clonal single-nucleotide variants (SNVs): 0.063 compared with 0.079 in the parental cell line (Figure 3.2 panel B). UVB-irradiated B2905 cells grew at a slower rate in vitro compared with non-irradiated B2905 cells (Figure S2E), the irradiated cell line gave rise to tumors with an increased growth rate when transplanted into immunocompetent syngeneic mice (Figure 3.2 panel C and D). This effect was not cell line specific because irradiated B16F10.9 cells showed the same pattern of reduced growth in vitro and increased tumorigenicity in vivo. We additionally assessed whether tumors derived from these two lines, parental B2905 and UVB-irradiated B2905, had a differential response to PD-1 blockade. We found that the response of mice with the UVB-irradiated cell line to anti-PD-1 treatment was considerably milder than the response of those with parental B2905 cells (Figure 3.3). Given that the UVB signature cannot predict checkpoint blockade response in melanoma patients [149] and that excessive TMB did not reduce tumor growth, we hypothesized that differences in heterogeneity may play a role in mediating tumor growth in vivo.

3.3.3 Increasing ITH leads to reduced T-cell reactivity to neo-antigens and T cell infiltration in-vivo

⋆⋆ Experimental analyses done by the Yardena Samuels' lab.

We next evaluated whether the growth rates of the tumors harboring dif-

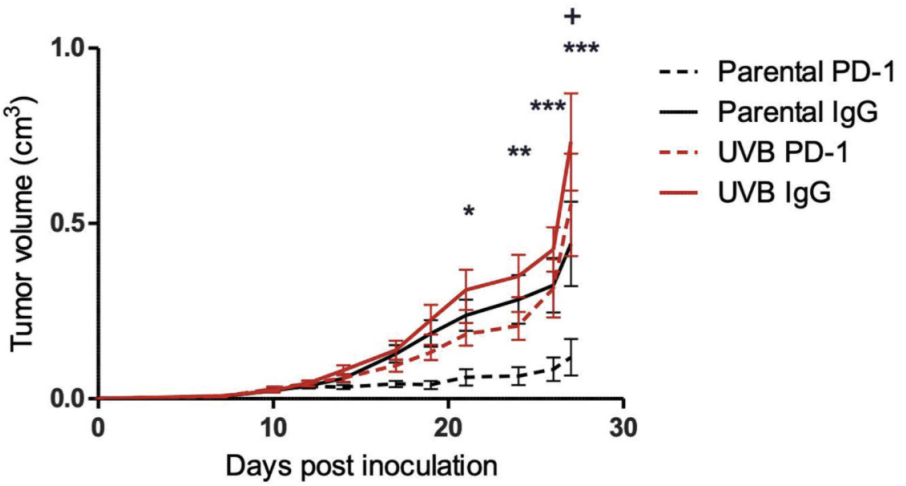

Figure 3.3: In vivo tumor growth in mice inoculated with parental B2905 (Black) or UVB irradiated B2905 (red) lines, treated with anti-PD-1 or IgG control antibodies at days 6, 9, and 12 post cells inoculation (n = 11-12). Data are mean $\pm SE$. Comparisons between parental B2905 tumors treated with IgG or anti-PD-1 treated are depicted by asterisks, whereas comprisons between UVB B2905 tumors treated with IgG or anti-PD-1 are depcited by cross. $\star p < 0.05, \star\star p < 0.01, \star\star\star p < 0.001$, one-way ANOVA followed by Tukey's post hoc test

ferent degrees of heterogeneity are mirrored by the degree of T cell reactivity in vivo. This was assessed by extracting total T cell receptor β+ (TCRβ+) TILs (tumor-infiltrating lymphocytes) from non-irradiated parental B2905 tumors, UVB-irradiated B2905-derived tumors, and B2905 SCC 2-derived tumors. To assess T cell reactivity, we measured the fraction of TILs positive for the cytotoxic mediator Granzyme B coupled with expression of CD107a, a degranulation marker [6, 198]. Although total Granzyme B+ fractions were similar in TILs derived from all tumors, the Granzyme B+ CD107a+ fraction of TILs was significantly reduced in UVB-irradiated B2905-derived tumors, whereas it remained similar in both the parental and the SCC 2-derived tumors (Figure 3.4, panel A). In addition, SCC 2-resident TILs contained a much higher interferon-+ fraction (Figure 3.4, panel B), indicating stronger TIL activation and cytotoxicity. To substantiate these results, we sorted

CD8+ TILs from UVB-derived and SCC 2-derived tumors 16 days after inoculation, performed RNA sequencing (RNA-seq), and analyzed the TILs for their CYT score. CD8+ TILs isolated from SCC 2-derived tumors had a higher CYT score, recapitulating the high CYT scores of the low-ITH TCGA melanoma patients (Figure 3.4 panel G and H). Furthermore, this score significantly correlated with tumor weight (Figure 3.4 panel C and D). Thus, the SCC 2-derived tumors, which had low ITH and were ultimately rejected in vivo, were more immunogenic than their parental heterogeneous UVB-irradiated, aggressive B2905-derived tumors, which had high ITH.

In addition to the immune composition of the tumor microenvironment, the spatial distribution of TILs within the malignant mass, in particularly immune infiltration into the tumor core, correlates with better survival and treatment success [112, 117]. Immunohistochemistry (IHC) and immunofluorescence analyses of tumor sections revealed that, although tumors derived from all three cell lines accumulated CD8+ TILs in the tumor margin, those derived from SCC 2 featured both higher penetration of CD8+ cells (Figure 3.4, panel E) and massive infiltration of TILs into the tumor core (Figure 3.4 panel F and G). We recapitulated these data in three additional SCCs that also formed tumors large enough for IHC analysis. We next quantified the levels of regulatory T cells (Tregs), which are known to suppress anti-tumor immunity and promote tumor growth [66, 225] by CD3+ Foxp3+ immunofluorescence (IF) staining of these tumors and found a direct correlation between ITH and Treg levels (Figure 3.4 panel H and I). In conclusion, low-ITH tumors show enhanced CD8+ T cell infiltration to the tumor core, a lower presence of immunosuppressive

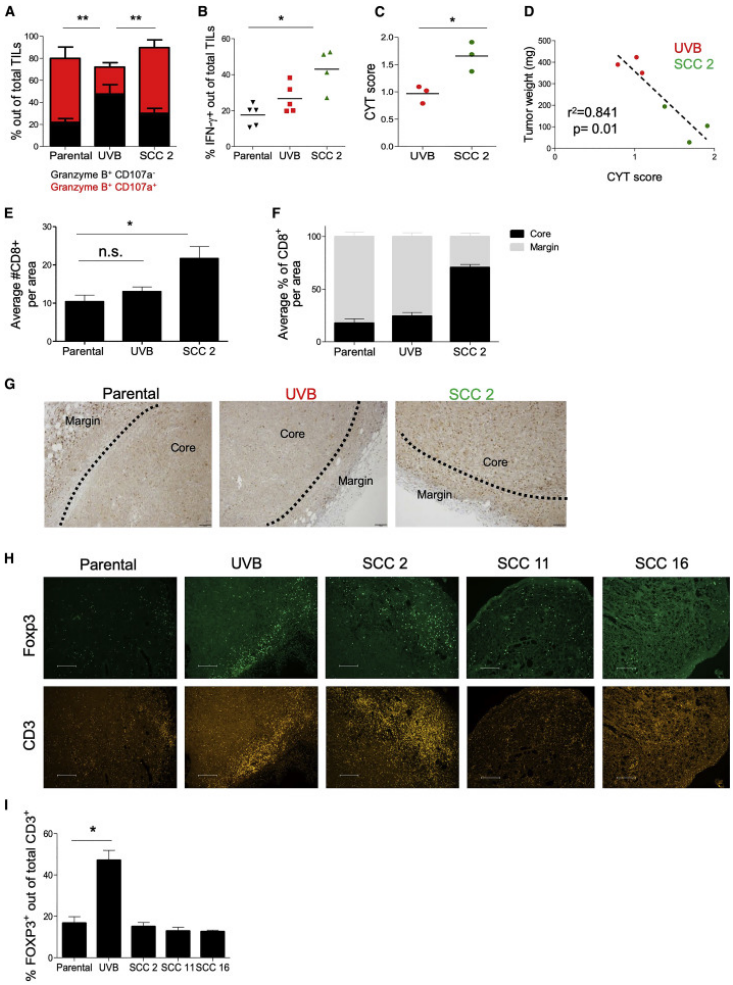

Figure 3.4: (A) Flow cytometry analysis of the Granzyme B and CD107a population in total TCRβ+ TILs on day 19. n = 4–5; data are mean $\pm SE$. $\star\star\, p < 0.01$ for Granzyme B$^+$ CD107a$^+$ TILs, two-way ANOVA followed by Bonferroni's post hoc test.(B) Flow cytometry analysis of interferon-gamma (IFN$-\gamma$) in total TILs on day 19. n = 4–5, $\star p < 0.05$, Kruskal-Wallis test followed by Dunn's multiple comparisons test. (C) CYT score derived from RNA-seq data of sorted CD8+ TILs from UVB-irradiated B2905 and SCC 2 tumors on day 15. $\star p < 0.05$, Mann-Whitney U test. (D) Pearson correlation between CYT score and weights of tumors in Figure 3C. (E) Quantitation of total CD8+ TILs in the indicated tumors. Four sections from each tumor and three tumors derived from each cell line were examined. A significant difference was observed between parental cells and SSC 2 but not between parental cells and UVB. Data are mean $\pm SE$. $\star p < 0.05$, one-way ANOVA followed by Tukey's post hoc test. (F) Relative quantitation of the average percentage of CD8+ TILs in the tumor core versus the margin of the tumors described in (E). Data are mean $\pm SE$. (G) Representative immunohistochemical stain for CD8 in slides taken from tumors derived by parental, UVB and SCC 2 on day 10 after cell inoculation. The scale bars represent 100 μM. (H) Immunofluorescence stains of CD3 and Foxp3 in tumors derived from B2905 parental, UVB, and SCC 2, 16, and 11 on days 1011 after cell inoculation. 3–4 sections from each tumor and two tumors derived from each cell line were examined. The scale bars represent 200 μM. (I) Relative quantitation of the percentage of Foxp3$^+$ of CD3$^+$ TILs described in (H). Data are mean $\pm SE$. $\star p < 0.05$, one-way ANOVA followed by Tukey's post hoc test.

Tregs, and higher degranulation and cytotoxicity compared with high-ITH tumors. This indicates that, indeed, low-ITH tumors elicit a strong anti-tumor response, whereas high ITH tumors are relatively non-immunogenic.

3.3.4 Systematic clone mixing experiments show that both the number of clones and their genetic diversity affect host immune rejection

⋆⋆ Experimental analyses done by the Yardena Samuels' lab

To further study the role of ITH in tumor rejection, we systematically generated tumors with defined states of heterogeneity using different combinations (mixtures) of the above-described 20 SCCs that were derived from the original, highly heterogeneous UVB-irradiated cell line (3.2). The individual SCCs were mixed in a controlled manner to dissect the functional ramifications of the two fundamental components of tumor heterogeneity: (1) the number of clones comprising the tumor and (2) the genetic diversity between them. To choose relevant clones for the mixing experiments, we performed a phylogenetic analysis of the heterogeneous UVB cell line. This yielded a phylogenetic tree with six terminal branches (TBs), numbered TB-4 to TB-10 (Figure 3.5 panel A). An almost identical clustering was obtained using an orthogonal analytical methodology. We then placed the 20 SCCs on the various terminal branches of the tree, based on their sequence similarity. To study the role of tumor diversity in determining tumor growth, we inoculated four different mixtures of 3 SCCs and monitored their growth, as shown in Figure 6B. To

achieve genetically diverse mixes, each mix contained clones from 3 different TBs of the UVB-irradiated phylogenetic tree (denoted as across branches [3AB]) (Figure 3.5 panel B and E). As seen in Figure 3.5, panel B, although diverse, none of the 3 clone mixes formed a large tumor, even 35 days after the mixes were inoculated. However, increasing the number of branches included in the mix from 3 to 6 (one clone from each TB [6AB]; Figure 3.5 panel E) results in significantly larger tumors (Figure 3.5 panel B and C) (group factor $p = 0.0238$ when 6AB is compared with the 3AB mixes by two-way ANOVA versus group factor $p = 0.1614$ when the 3AB mixes are compared without 6AB). Doubling the number of clones included from each of the six TBs (two clones from each TB [12AB]) further increased the subclonal/clonal mutation ratio (Figure 3.5 panel D) and produced even more aggressive tumors (Figure 3.5 panel C).

We next evaluated the functional effects of the tumor's genetic diversity while controlling for the overall mutational load. To this end, we compared the growth of tumors generated from a mixture of clones originating from a single TB (6 SCCs within TB-4 [6WB]) with that of those generated from the 6AB mix described above (comprising clones from TB-4, TB-6, TB-7, TB-8, TB-9, and TB-10) (Figure 3.5 panel C and E). These two mixes have the same number of clones (six) and approximately the same mutational loads (Figure 3.5 panel F) but vary in their genetic diversity, as assessed by their clonal versus subclonal mutation ratios (Figure 3.5 panel D). Despite having similar mutational loads, we identified striking differences in growth between 6AB and 6WB (Figure 3.5, panel C). Similarly, we next compared the tumor growth curves of a mixture of 12 SCCs derived from one branch

(12WB, derived from branch 5, which contains TB-4 and TB-7) with the growth of the 12AB mix (two SCCs from TB-4, TB-6, TB-7, TB-8, TB-9, and TB-10) (Figure 3.5 panel C and E). Again, there were clear differences in tumor growth between 12AB and 12WB. Moreover, even though 12AB had a higher mutational load than 6AB (Figure 3.5 panel D), its growth surpassed that of 6AB. This indicates that an increased mutational load is not sufficient to drive tumor rejection. The 12AB tumors were still not as aggressive as the UVB irradiation-derived tumors (Figure 3.5 panel D), emphasizing that the latter tumors harbor a higher degree of ITH. Taken together, these results testify that both the number of tumor subclones and their genetic diversity play important roles in mediating tumor growth and rejection.

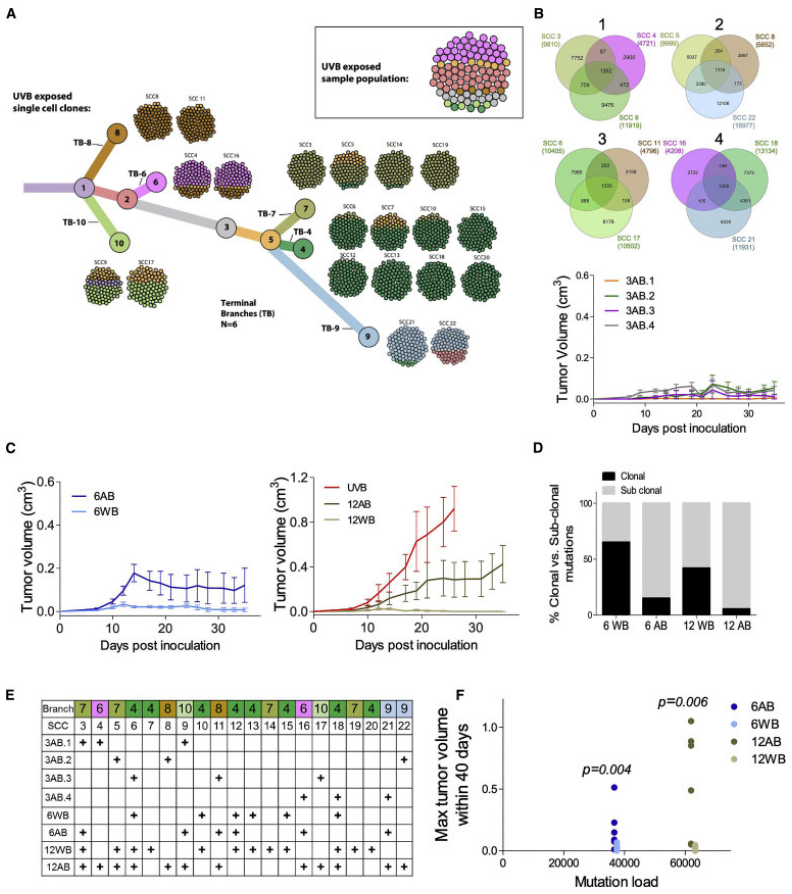

Figure 3.5: (A) Phylogenetic tree representation of the UVB-irradiated B2905 cell line. The tree depicts the results from mutation clustering analysis, which was used to define the distinct subclones present within the UVB cell line. The phylogenetic relationship between subclones is shown, and then each of the 20 UVB derived SCCs is mapped onto the subclonal branch with the highest genetic similarity. Each of the 20 SCCs is depicted as a ball of 100 tumor cells, with the color coding reflecting the percentage frequency of each branch in each SCC sample. Shown in the top right box is a representation of the UVB parental sample, again shown as a ball of 100 tumor cells, color-coded to match the subclonal branches. (B) Top: Venn diagrams for the four 3AB mixes inoculated, representing the number of protein-coding mutations and their intersections between the SCC in each mix. Bottom: in vivo tumor growth curves of the four different 3AB mixes. n = 5. Data are mean $\pm SE$.(C) Left: in vivo tumor growth curves of the 6WB mix (within TB-4) and 6AB mix (one SCC from each TB). n = 4–5. Right: in vivo tumor growth curves of the 12WB mix (within TB-5) and 12AB mix (two SCC from each TB) and the UVB-irradiated B2905 cell line. n = 5–6. Data are mean $\pm SE$. (D) Percent clonal versus sub-clonal mutations in the mixes described in (C). (E) The SCC included in each mix described in (B) and (C). (F) The association between the 6AB, 6AB, 12AB, and 12WB mix mutation number (unique) and the maximal tumor volume size (cubic centimeters) within 40 days. Each dot represents an individual mouse. The graph shows statistical significance between the 6 and 12 mixes but not between mutation number and tumor volume (Wilcoxon rank-sum test)

3.3.5 Tumor clonal diversity predicts responses to immune checkpoint blockade therapy even after controlling for tumor mutation burden

⋆⋆ Analysis done in collaboration with Kevin Litchfield

To further evaluate the extent to which the number of clones and their genetic diversity affect the anti-tumor immune response in human data, we analyzed four previously published melanoma checkpoint inhibitor cohorts from Snyder et al., 2014 [217], Riaz et al., 2017 [185], Hugo et al., 2016 [99], and Van Allen et al. (2015) [233]. Given the results of the mixing experiment that show that both the number of clones and their diversity are important determinants of tumor growth, we analyzed patient data using the Shannon diversity index (SDI), a formal diversity metric that quantitatively measures both the number of clones and the diversity of the mutations across clones in one index. As an example, a tumor with a low SDI would have nearly all of its mutations concentrated in just one clone (a large truncal neoantigen burden). In contrast, a high-SDI tumor would have a high number of clones with mutations spread evenly or diversely across each clone (a large branched neoantigen burden) (Figure 3.6 panel A). The first cohort analyzed (Snyder et al., 2014) comprised data from 54 patients treated with anti-CTLA-4 therapy. We found its SDI index to significantly associate with overall survival ($p = 0.0064$, SDI tested as a continuous variable, z-test from the Cox proportional hazard model; Figures 3.6 panel B and F). Patients with a higher diversity tumor (as measured by

SDI) had poorer survival, with a hazard ratio (HR) of 8.8 (95% confidence interval, 1.8–41.6) per unit increase in SDI (Figure 3.6 panel F). In the second cohort (Riaz et al., 2017), containing 57 patients treated with anti-PD1 therapy, we observed a comparable but non-significant pattern ($p = 0.079$, HR = 2.2 [0.9–5.5] per unit increase in SDI (Figure 3.6 panel C and F). In the third cohort (Hugo et al., 2016), composed of 21 patients treated with anti-PD1 therapy, again a comparable but non-significant pattern was noted ($p = 0.096$, HR = 4.2 [0.8–23.8] per unit increase in SDI; Figures 7D and 7E). In the final cohort, which had data available from 70 patients treated with anti-CTLA4 therapy, no significant association between SDI and overall survival was detected (Figures 7E and 7F); it should be noted, however, that this result is consistent with previous ITH analyses in this cohort (McGranahan et al., 2016) and may be explained by the high level of pre-treatment in this cohort, making biomarker analyses more challenging. Given that all four datasets are of fairly limited size, we performed a meta-analysis across all four studies, which yielded an overall significance value of Pmeta = 0.0105, testifying that clone number and genetic diversity between clones are drivers of the immunotherapy response in human cohorts. Importantly, this result remained significant after adjusting for tumor purity in a multi-variable analysis for each cohort, with updated Pmeta = 0.012 for the SDI variable (across all four studies), and Pmeta = 0.15 for tumor purity, suggesting that the latter is not a confounding variable in our analysis. Similarly, we corrected for TMB in the multi-variable analysis for each cohort, which yielded an updated Pmeta = 0.039 for the SDI variable and Pmeta = 0.33 for TMB.

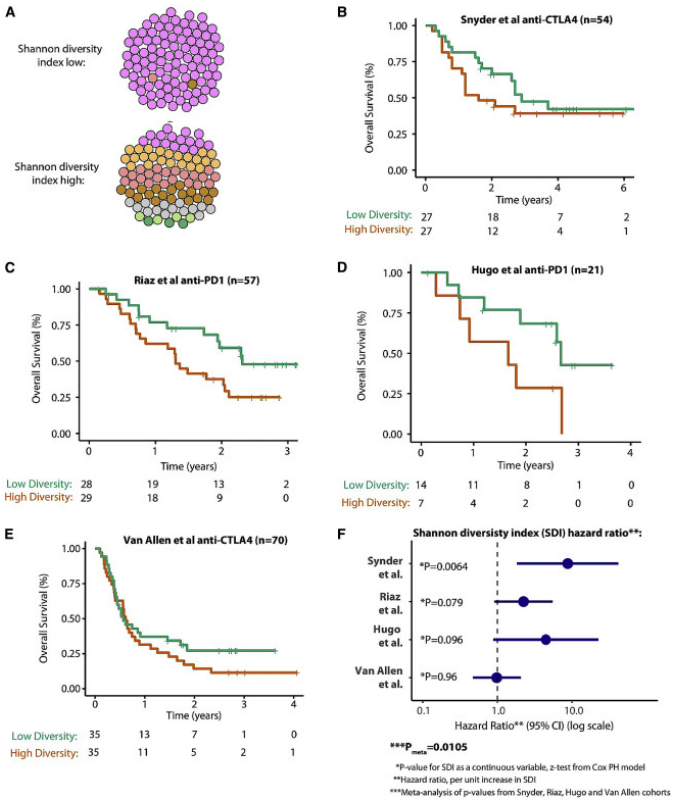

Figure 3.6: (A) The cartoon illustrates two examples of the SDI, top low SDI (the tumor is predominantly composed of one major clone) and bottom high SDI (the tumor is composed of multiple clones with higher evenness between clones). SDI is measured using individual tumor subclones (from Pyclone clustering) as types and the somatic mutations as entities so that a tumor with a low SDI would have nearly all mutations concentrated in just one clone, and, in contrast, a tumor with a high SDI would have a higher number of clones, with mutations spread evenly or diversely across each clone. (B) The SDI analysis applied to the Snyder et al. (2014) anti-CTLA4 dataset [217]. Overall survival Kaplan-Meier plots are shown for with patients with a high SDI in red (SDI above median value in cohort) and a low SDI in green. The number of patients at risk by time point is shown in the table below. (C–E) The same data format as in (B) for the Riaz et al. (2017) anti-PD-1 dataset [185] (C), Hugo et al. (2016) anti-PD-1 dataset [99](D), and Van Allen et al. (2015) anti-CTLA4 dataset [233] (D), respectively. (F) Forest plot showing the HR for the SDI in each dataset, with the HR value corresponding to the survival risk per unit increase (i.e., each +1 increment) in the SDI. For significance analysis, SDI is tested as a continuous variable (to show a continuous association across the full range of data) using a Cox proportional hazard model (other clinical predictors, e.g., stage, are not included).

3.4 Discussion

Here we have established a framework that enables one to tease apart and study the effect of TMB and ITH on tumor aggressiveness, evaluating their influ-

ence on anti-tumor immunity in a controlled manner. Our findings in mice suggest that, in melanoma, an essential genetic determinant of anti-tumor immune response is tumor heterogeneity. These results corroborate previous reports that clonal neoantigens are associated with a more robust tumor infiltrate and clinical outcome, with and without checkpoint inhibitor blockade [143]. By systematically generating tumors composed of different SCC mixes in a designed, controlled manner, we further dissected the two major components of a tumor's ITH, finding that both the number of distinct clones composing the tumor and the degree of their genetic diversity influence tumor aggressiveness.

Our experimental mouse data are mirrored in TCGA melanoma patients, where the overall survival rate is significantly higher in tumors with a fewer number of clones, and the combination of number of clones and diversity (their SDI) is inversely associated with overall survival in immune checkpoint inhibitor-treated cohorts. These findings, which tightly match our experimental findings in mice, further support the detrimental influence of tumor heterogeneity on the anti-tumor immune response in humans, in keeping with previous studies [143].

Alongside the effects on tumor growth and responsiveness, the complex mechanisms behind the modulation of anti-tumor immunity by tumor heterogeneity need to be further addressed in future studies. We suggest that diminishing tumor heterogeneity exposes tumor cells by reducing their neoantigen landscape, bringing reactive neoantigens to the "frontline," thus better exposing them to immune detection. This, in turn, leads to enhanced infiltration into the tumor core, elevated effector cytokines, and heightened degranulation. When neoantigen-specific CD8+

T cells are able to infiltrate and kill tumor cells, more tumor antigens are exposed to the tumor microenvironment, further promoting neoantigen uptake and presentation by tumor-associated DCs, enhancing the ability of the immune system to reject the tumor. In contrast, in more heterogeneous tumor cell populations, tumor cells could have a better chance of escaping immune surveillance because the reactive neoantigens undergo "dilution" within the tumor relative to other neoantigens. The total outcome is weaker anti-tumor immunity, manifested by reduced immune infiltration into the tumor core and dampening of TIL degranulation, cytotoxicity, effector cytokine secretion, and proliferation. In addition to CD8+ T cells, we found lower numbers of Tregs in tumors derived from SCCs (low ITH) than in UVB-derived, more heterogeneous tumors, indicating a strong immunosuppressive tumor microenvironment in high-ITH tumors that is resolved in single-cell-derived tumors. Overall, our results are consistent with the recent hypothesis by Gejman et al. (2018) [71] that, because of increased antigenic variability, the relative expression of each neoantigen is lowered in tumors with increased ITH, diminishing the TILs' ability to home to their target cells and mount a sufficient cytotoxic response.

In addition to the differential infiltration of CD8+ T cells and differential Treg accumulation observed in the tumors, other immune mechanisms also likely play a part in the reduced response to heterogeneous versus homogeneous tumors. These may involve non-Treg CD4+ cells, which are important for priming of CD8+ T cells [27] and recognition of MHC class II-borne tumor antigens [256]. Different CD4+ T cells effector subsets can have direct or indirect anti-tumor immunity. These subsets include CD4+ cytotoxic T cells that can directly eliminate MHC class II+ tumors

[90] and CD4+ Th1 and Th17 cells that can mediate elimination of tumor cells in an antigen-specific manner [158, 180]. Indeed, strong anti-tumor responses of CD4+ cells against tumor MHC class II neoantigens in cancer patients have been reported [130, 230, 236]. This suggests an additional level of complexity within the tumor-immune interface and a significant clinical potential for future therapies.

Additional immune subsets other than T cells that may play a role in this setting encompass M1 and M2 macrophage polarization [33], NK cells [88], DCs [58], or neutrophils [49]. To fully elucidate the immune profiles of ITH-high versus -low tumors, cutting-edge, high-dimensional techniques such as single-cell RNA-seq [127, 201] and CyTOF [86] and state-of-the-art analysis algorithms such as CIBERSORT [164] could be utilized in follow-up studies.

Although we show that high ITH impairs the immune system response, tumors with impaired immune responses can likely still acquire high levels of ITH. Thus, impaired immune response and ITH levels are tightly associated. However, whether ITH is a cause or a consequence of tumor progression or both is not fully elucidated. Interestingly, previous studies have shown that functional cooperation between genetically distinct subclones can be essential for overcoming environmental constraints and, thus, affect tumor maintenance and growth [48, 140] and metastatic behavior [102]. Of note, it has been shown recently that the immune system as well as checkpoint immunotherapy can select for low-ITH tumors [152]. Understanding the complex interactions between tumor heterogeneity and the immune response and how they change during tumor evolution still remains a challenge.

Despite the strengths, there are several shortcomings of our study. Specifically,

the single-cell cloning process inherently involves in vitro selection of clones. This may miss clones with a low survival capability in vitro, which does not necessarily reflect their functional importance in vivo. Likewise, we acknowledge the limitations of accurately assessing ITH from a single biopsy sample in the TCGA and immune checkpoint blockade (ICB) datasets because of the narrow sampling frame of taking just one sample from one spatial location. The sequencing depth, tumor purity, choice of processing pipeline, and nature of the single biopsy (primary versus metastatic) may also affect ITH assessment, making it challenging to derive a single prognostic measure of ITH. We believe that additional studies that quantify ITH in large-scale cohorts with multi-region biopsies are likely to shed further light on the prognostic role of tumor ITH, providing a higher-resolution view of the fundamental trends outlined in this study.

In summary, our findings show the value of evaluating ITH as an important determinant of melanoma patients' response to checkpoint therapy. They also support the notion that clonal neoantigens are more likely to lead to better cancer vaccines [143, 202]. On the flip side, our results cast doubt on the notion that excessive mutagenesis, directed to enhance TMB, can enhance the efficacy of immunotherapy. Indeed, it is conceivable that excessive neoantigen heterogeneity may actively impair a productive anti-tumor immune response. In conclusion, our functional data support recent findings that the clonality of a tumor can be used as a biomarker for predicting better outcomes in melanoma and may improve patient matching to current immunotherapy in a manner complementary to mutational load. We suggest that ITH is a strong determinant of immune response and immunotherapy success

in melanoma, highlighting the potential importance of assessing it in the clinic. The github source code related to the TCGA analysis done in this work is available at: `https://github.com/spatkar94/UVB_Melanoma.git`

Chapter 4

Factors driving acquisition of cancer type-specific chromosomal aneuploidies

⋆⋆ This work was done in collaboration with Dr. Noam Auslander and currently under review at Genome Medicine.

4.1 Overview

In solid tumors of epithelial origin, i.e., carcinomas, and in certain other solid tumors such as glioblastoma multiforme and malignant melanoma, aneuploidies of specific chromosomes define the landscape of somatically acquired genetic changes [116, 115, 161, 93, 187]. In fact, aneuploidy is present in about 90% of solid tumors [21]. Remarkably, the distribution of ensuing genomic imbalances is cancer-type specific [93, 188]. For instance, colorectal carcinomas are defined by extra copies of chromosomes and chromosome arms 7, 8q, 13q and 20q, accompanied by losses of 8p, 17p and 18q [190]. In contrast, cervical carcinomas invariably carry gains of chromosome arms 1q and 3q. In other words, a gain of 3q is not observed in colorectal cancer, and cervical carcinomas do not have copy number gains of, e.g., chromosomes 7 or 13q [93, 187, 188]. Furthermore, cancer-type specific chromosomal aneuploidies emerge in dysplastic, i.e., not yet malignant, lesions, that are prone to progress to invasive disease [190, 178, 22, 226, 55]. Numerous cancer-type specific aneuploidies

originate at early stages of tumorigenesis, yet are retained in late stage tumors and in metastases, as reflected in the TCGA database [178]. The cancer-type specific distribution of genomic imbalances was recently confirmed in two comprehensive pan-cancer analyses of several thousand tumors [22, 226]. Although some intra-tissue differences can be observed for certain tumor subtypes arising from the same tissue, different tumor types from the same tissue tend to cluster together (e.g., low-grade gliomas cluster with glioblastomas as do clear cell and papillary renal cell carcinomas) . On one hand, it is possible that loss or gain of particular chromosomes or their fragments during carcinogenesis target the gain of specific oncogenes or the loss of tumor suppressors located on these chromosomes [21, 55, 15]. On the other hand, it is well known that chromosome-wide alterations of gene expression levels follow genomic copy number changes [232, 189], i.e., the transcripts of genes that are located on gained chromosomes are more, and those on lost chromosomes are less abundant. This correlation has been firmly established in primary human carcinomas, in derived cell lines, and in experimental cancer models [232, 247, 231, 221, 62, 191]. Hence the gain or loss of specific chromosomes can potentially act as a mechanism to maintain tissue specific gene dosage. Given this background, we decided to explore how the frequencies of chromosomal arm gains and losses in specific cancer types correlate with (i) mean chromosome arm gene expression levels of their normal tissue of origin, and (ii) the chromosomal distribution of previously identified or newly implicated tissue specific driver genes. Our exploratory analysis unearths a complex picture of factors shaping the evolution of tumor karyotypes in which recurrent chromosomal alterations can potentially "hardwire" expected

chromosome-wide gene expression levels of their normal tissue of origin in addition to targeting tissue-specific driver genes.

4.2 Methods

4.2.1 Tissue and tumor type inclusion

Initially, all 33 TCGA tumor types were considered for analysis. Chromosome arm-wide gain and loss data data for 33 tumor types were obtained from Taylor et al. [226], cancer gene expression data and normal gene expression was obtained from TCGA (`xenabrowser.net`) and GTEx (GTEx analysis V6p), respectively, using Reads Per Kilobase of transcript, per Million mapped reads (RPKM) values with no additional normalization. The RPKM values are already library size normalized, through dividing by the total number of reads in a sample, therefore accounting for whole genome doubling events. Moreover, because the GTEx samples are of healthy individuals, it is unlikely that any of these samples harbor whole genome doubling events. Processed methylation datasets of normal tissues were collected from the Gene Expression Omnibus (GEO) database. For consistency, we restricted our search to datasets where methylation was quantified using the same platform (Illumina 450K). This approach resulted in the identification of 18 tissue specific methylation datasets, which were analyzed together. For analysis comparing tumor and normal tissues, tumor samples from 25 tumor types and 19 corresponding normal GTEx tissues of origin were considered (See Table 4.1). Likewise, for comparing tissue specific methylation and expression levels, only 11 tissues which had

a matching methylation dataset available were considered.

Table 4.1: Cancer type-normal tissue pairs evaluated

Cancer type (TCGA)	normal tissue of origin with available gene expression (GTEx)	normal methylation data available
ACC	Adrenal Gland	No
BLCA	Bladder	No
LAML	Blood	No
LGG	Brain	Yes
GBM	Brain	Yes
BRCA	Breast	No
CESC	Cervix Uteri	Yes
COAD	Colon	Yes
READ	Colon	Yes
ESCA	Esophagus	Yes
KIRP	Kidney	Yes
KIRC	Kidney	Yes
KICH	Kidney	Yes
LIHC	Liver	Yes
LUSC	Lung	Yes
LUAD	Lung	Yes
OV	Ovary	Yes
PAAD	Pancreas	Yes
PRAD	Prostate	Yes
SKCM	Skin	Yes
STAD	Stomach	No
TGCT	Testis	No
THCA	Thyroid	No
UCEC	Uterus	Yes
UCS	Uterus	Yes
DLBC	NA	No
THYM	NA	No
MESO	NA	No
CHOL	NA	No
HNSCC	NA	No
UVM	NA	No
PCPG	NA	No
SARC	NA	No

4.2.2 Curation and pre-processing normal tissue specific methylation datasets

We curated a list of 18 Illumina 450K methylation datasets covering 11 organ tissues from GEO (See Table 4.2). These were datasets spanning different studies comparing methylation levels of organ tissues between diseased and normal control individuals. We only selected methylation profiles of normal control individuals for further analysis. Moreover, multiple datasets containing samples coming from the same organ tissue were merged to generate one methylation dataset per organ. The methylation data of each dataset was pre-processed in the following steps:

- Filtering out probes within 15 base pairs of single nucleotide polymorphisms [60].
- Re-normalizing the beta values between type 1 and type 2 probes using beta mixture quantile normalization [228]. This minimizes biases that may arise due to sensitivity differences between the two probe designs.

Table 4.2: list of curated methylation datasets

Dataset ID	Tissue
GSE32146	Colon
GSE40360	Brain, frontal lobe
GSE61107	Brain, frontal cortex
GSE88890	Brain, cortex
GSE89702	Brain, cerebellum
GSE89703	Brain, hippocampus
GSE89705	Brain, Striatum
GSE62640	Pancreas
GSE51954	Skin
GSE90124	Skin
GSE52401	Lung
GSE61258	Liver
GSE61446	Liver
GSE51820	Ovary
GSE46306	Cervix
GSE45187	Uterus
GSE52826	Esophagus
GSE59157	Kidney

4.2.3 Computation of the chromosome arm imbalance score in cancerous tissues

We used the TCGA sample-wise chromosomal arm gain and loss data provided by Taylor et al. [11], where the ploidy was determined via the ABSOLUTE algorithm [38]. Independent chromosome arm copy number alterations were distinguished from whole genome duplication events by comparing the absolute integer copy number of chromosomal arm regions to the baseline tumor ploidy. Each segment was designated

as gained, deleted, or neutral compared to the ploidy of each sample. The scores of each arm are -1 if lost, +1 if gained, 0 if non-aneuploid, and "NA" otherwise. For sake of consistency, all "NA" entries were re-set to 0 (i.e, we consider those samples non-aneuploid for that arm). The discrete representation was used because it is most fitting to describe arm-level changes, which may be either gained (1) or lost (-1) by definition, rather than continuous GISTIC data, which is better suited for studying targeted focal copy number alterations. For each of the 39 chromosomal arms we define an *arm imbalance score* for a set of cancer types sharing the same tissue of origin (or a singular cancer type), by computing the difference between the frequency of gains and losses. Formally:

$$\text{Arm Imbalance Score}(A_i, T_j) = \frac{\sum_{\text{samples s} \in T_j} I_{sG}(A_i) - \sum_{\text{samples s} \in T_j} I_{sL}(A_i)}{\text{Number of samples in } T_j} \tag{4.1}$$

Where A_i is chromosomal arm i (of 1 to 39 chromosomal arms), T_j is the tissue of origin of all tumor types arising from tissue j and the indicators $I_{sG}(A_i)$ and $I_{sL}(A_i)$ are defined as:

$$I_{sG}(A_i) = \begin{cases} 1 & \text{if sample } s \text{ has a gain of arm } A_i \\ 0 & \text{otherwise} \end{cases}$$

$$I_{sL}(A_i) = \begin{cases} 1 & \text{if sample } s \text{ has a loss of arm } A_i \\ 0 & \text{otherwise} \end{cases}$$

Hence, arms that are more frequently gained are assigned positive scores, while

arms that are more frequently lost are assigned negative scores. Arms that are neither gained nor lost, and arms where the frequency of gains and losses is comparable are assigned neutral (≈zero) score. However, the latter is negligible since chromosome arms that are frequently gained are rarely lost in a specific tumor type and vice versa. This score is hence equivalent to the mean value of gains/loss incidences in set of tumor types considered and chromosomal arm.

4.2.4 Permutation tests to evaluate correlation significance

In this study, we compute correlations across cancer/tissue types, and across chromosomal arms. To evaluate whether the magnitude of correlations is significant compared to random, we employ a permutation test, to estimate a background null distribution of the number of positive correlations. We therefore repeat 1000 iterations of randomly shuffling the cancer/tissue pairing and 1000 iterations of randomly shuffling the arm-level pairing. We compare the number of positive correlations P, achieved with the true pairings to this background (N_i,$i = 1, 2, \ldots, 1000$), to compute a p-value and accept or reject the null hypothesis, denoted as $\frac{\sum_{i=1}^{1000} N_i > P}{1000}$. In a similar manner, we test whether mean arm-wide gene expression levels of each of the 39 chromosome arms in a sample are informative for predicting the sample's tissue of origin, compared to the background of any random aggregation of gene expression into 39 groups. Therefore, we design a permutation test with 1000 iterations. In each iteration, we quantify how accurately we can predict tissue of origin based on randomly aggregating genes into 39 groups with similar sizes as that of

chromosomal arm assignment. We evaluate the number of times (out of 1000) in which the multiclass prediction accuracies of the shuffled predictor (N_i, with random aggregation of genes into 39 groups) exceeded the original predictor (P, with the aggregation of genes to 39 groups by chromosomal arm), to derive an empirical permutation p-value, denoted as $\frac{\sum_{i=1}^{1000} N_i > P}{1000}$

4.2.5 Quantile normalization of gene expression and methylation values for cross tissue comparison and visualization

To enable side-by-side comparison and visualization of the arm imbalace scores with mean chromosomal arm mean gene expression levels in different normal tissues (and likewise in different cancers), the gene expression and arm-imbalance values need to be on the same scale. Hence, we additionally quantile-normalized the mean gene expression levels using the chromosomal arm imbalance distribution as reference, to enable visualization by generating similar expression distribution across different tissues. We applied the same approach to quantile normalize chromosome arm-wide mean methylation levels in normal tissues to visualize normal methylation against normal gene expression in each tissue.

4.2.6 Curation of chromosome-wide distribution of relevant oncogenes and tumor suppressors in each cancer type

We obtained a comprehensive list of known (or potential) oncogenes and tumor-suppressors driving each cancer type from a recent pan-cancer study con-

ducted by Bailey et al. [13]. This list was obtained from supervised machine learning predictions based on features derived from mutation, copy number, gene expression and mythylation changes observed in genes across different cancer types. Given a cancer type, the oncogenes-tumor suppressor imbalance score for each arm in a given cancer type (or collection of cancer types) is formally defined as follows: Oncogene-tumor suppressor imbalance score = fraction of driver genes on the arm that are oncogenes − the fraction of driver genes on the arm that are tumor-suppressors.

4.2.7 Normal and cancer tissue of origin classification and clustering

We classify normal (and likewise, cancer) samples using the chromosomal-arm level expression of those samples. For each sample , we calculate the mean gene expression level of the genes in each chromosomal arm. This results in 39 unique features per sample (one per arm). We then perform K-Nearest-Neigbors (KNN, with K=5, the value for which the best performance was observed for cancer type classification from K=3,5,7) classification with a Leave-One-Out cross validation (LOOCV), aiming to classify each sample based on the 39 arm level features, and calculate the resulting accuracy (percentage of correctly classified samples in the LOOCV). An analogous approach is taken for classification of tissue of origin based on methylation data. Additionally, to rule out potential confounding batch effects in gene expression data and the leave one out cross-validation procedure used, we re-estimate overall KNN performance using 5-fold cross validation. For performing hierarchical clustering of different tissue-types, each tissue-type is summarized as

a vector of 39 features; one for each arm. 4 different hierarchical clustering analyses are performed. For each one, a different set of 39 features is used. They are systematically listed below:

- Chromosomal arm imbalance score computed across all cancer types originating from the same tissue
- Mean arm-wide normal gene expression across all genes and all normal samples belonging to the same tissue.
- Mean arm-wide cancer gene expression across all genes and all samples originating from the same tissue
- Arm level oncogene-tumor suppressor imbalance score across all cancer types originating from the same tissue

4.3 Results

4.3.1 Chromosome arm imbalance scores of cancer types and mean chromosome arm-wide gene expression levels of their normal tissue of origin

Taylor and colleagues [226] comprehensively recorded for each tumor sample in the TCGA if a specific chromosome arm was gained or lost (while accounting for the baseline tumor ploidy).We used this data to compute the mean chromosome arm imbalance score of each arm in a given cancer type (or collection of cancer

types) emerging from the same tissue of origin. In short, this score measures the difference between the frequency of gains and losses of a specific chromosome arm. As a first step, we validated previous observations by showing that the mean gene expression levels over all genes and all samples from the same chromosome arms and cancer type included in the TCGA database, respectively, positively correlate with the corresponding arm imbalance scores (Figure 4.1, panel A). This analysis confirmed that genomic copy number alterations in cancer genomes directly affect gene expression levels. After having validated this correlation, we next computed the mean expression levels over all genes and all samples from the same chromosome arm and normal tissue, respectively, from the GTEx database. These values were then correlated with the mean chromosome arm imbalance scores of respective cancer types emerging from that tissue. Figure 4.1 panel B plots a heatmap with rows indicating chromosome arms. The chromosome arm wide mean expression levels in each normal tissue and corresponding arm imbalance scores in associated cancer types are juxtaposed and quantile normalized to the same scale for visualization and comparison.

In general, chromosome arms that are most frequently altered are either frequently gained or lost in each cancer type, with some notable exceptions (See for eg: chromosome 13q in gastrointestinal tumors). Nevertheless, the frequencies of these gains and losses vary by tissue of origin and result in varying arm imbalance scores across cancer types. Among the frequently altered chromosome arms, we see that chromosome arms 13q, 18q, 10q and 2p have the strongest correlations between their normal tissue specific mean expression levels and arm imbalance

scores, and these correlations are positive. When looking at each tissue individually (columns of Figure 4.1 panel B), we see the strongest correlations between the normal chromosome-wide mean expression levels and arm imbalance scores for brain, colon and kidney tissues and these correlations are also positive. Although the statistical power to assess the significance of these individual correlations is limited, we see that a majority of correlations (both at tissue and arm level) are positive. We evaluate the overall probability of getting so many positive correlations (both at the arm and tissue level), using a permutation test. To this end, we repeat 1000 times of randomly shuffling the chromosomal arm assignments (rows of Figure 4.1 panel B) and another 1000 for the tissue assignments (columns of Figure 4.1 panel B). We find that similar or higher correlations were found for the shuffled data in less that 5% of the cases, yielding a permutation $P < 0.05$ for both arm-wise and tissue-wise correlations. We additionally repeated this analysis for early stage tumors from the TCGA database (defined as tumors with AJCC stage classification of 0 or 1). Although the number of tumors available for analysis was further reduced, a similar trend of weak, but predominantly positive correlations was observed.

If certain chromosome arm aneuploidies might "hard-wire" the chromosome arm wide gene expression levels specific to their normal tissues, this suggests that one should be able to classify tissue of origin of normal and cancer tissue samples just based on the mean chromosome arm-wide gene expression levels of each of the 39 arms. To test this hypothesis, we obtained the mean gene expression levels for each arm in each normal tissue sample in GTEx (and likewise for each cancer sample in TCGA) resulting in 39 unique features. Then K-Nearest Neighbors (K-NN)

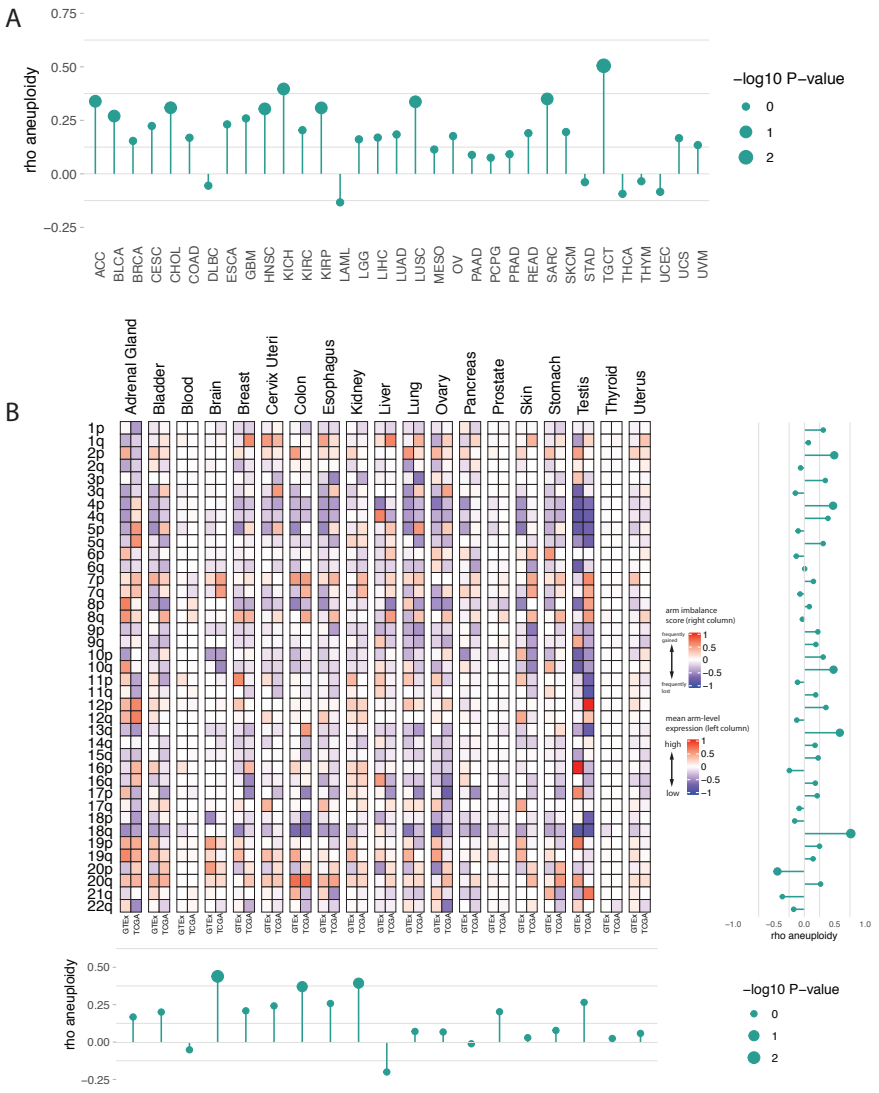

Figure 4.1: Correlations of chromosome arm-wide gene expression levels and chromosome arm-wide aneuploidies (A) Spearmn correlation plot of chromosome arm-wide gene expression levels in cancers and patterns of chromosome arm-wide gains and losses in cancers reported in the TCGA database. Bar plot represent the spearman rank correlations for each cancer type independently. The height of the bar reflects the correlation coefficient, and the size of the circle the significance. Size of 2 indicates p-value < 0.01, size of 1 indicates p-value < 0.1 and size of 0 indicates p-value < 1. (B) Spearmn correlation of chromosome arm-wide gene expression levels based on the GTEx database (left column) with chromosome arm wide aneuploidies in associated cancer types based on data reported in the TCGA database (right column), respectively, for 19 tissue entities. The arm imbalance score is reflected in colors: Red indicates more frequent gains compared to losses; blue indicates more frequent losses compared to gains. The hue of the colors indicates the frequency of copy number changes and the quantile normalized levels of mean chromosome arm-wide gene expression, respectively. Barplots shown beside each heatmap are the spearman rank correlations (horizontal bars indicate comparisons for each arm independently, vertical bars indicate comparisons for each tissue independently). A size of 2 indicates p-value < 0.01, a size of 1 indicates p-value < 0.1 and size of 0 indicates p-values < 1.

multi-class classification was applied with leave-one-out cross validation. We find that mean chromosome arm-wide gene expression can effectively classify the tissue of origin of both normal and cancer samples from GTEx and TCGA, respectively, and that the performance is generally better for normal tissues (Figure 4.2 panel A). The resulting accuracy was better for tissues with higher case numbers, as expected for KNN analyses. Furthermore, these results could never be obtained when the chromosome assignment of genes was randomly shuffled (by repeating 1000 shuffling of the chromosomal assignments of genes, empirical P-value < 0.001). A five-fold cross validation analysis yielded similar results. To visualize these classifications, we used t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction of the 39 dimensional feature space. We found that samples from the same normal tissues cluster closely in most cases (Figure 4.2 panel B), but to a lesser extent, for cancer entities (Figure 4.2 panel C). The separate sub-clusters within each tissue correspond to the different anatomical regions of the tissues that were sampled from GTEx. Overall, these results suggest that certain chromosomal aneuploidies acquired by tumors might hardwire expected tissue-specific gene expression levels of their tissue of origin.

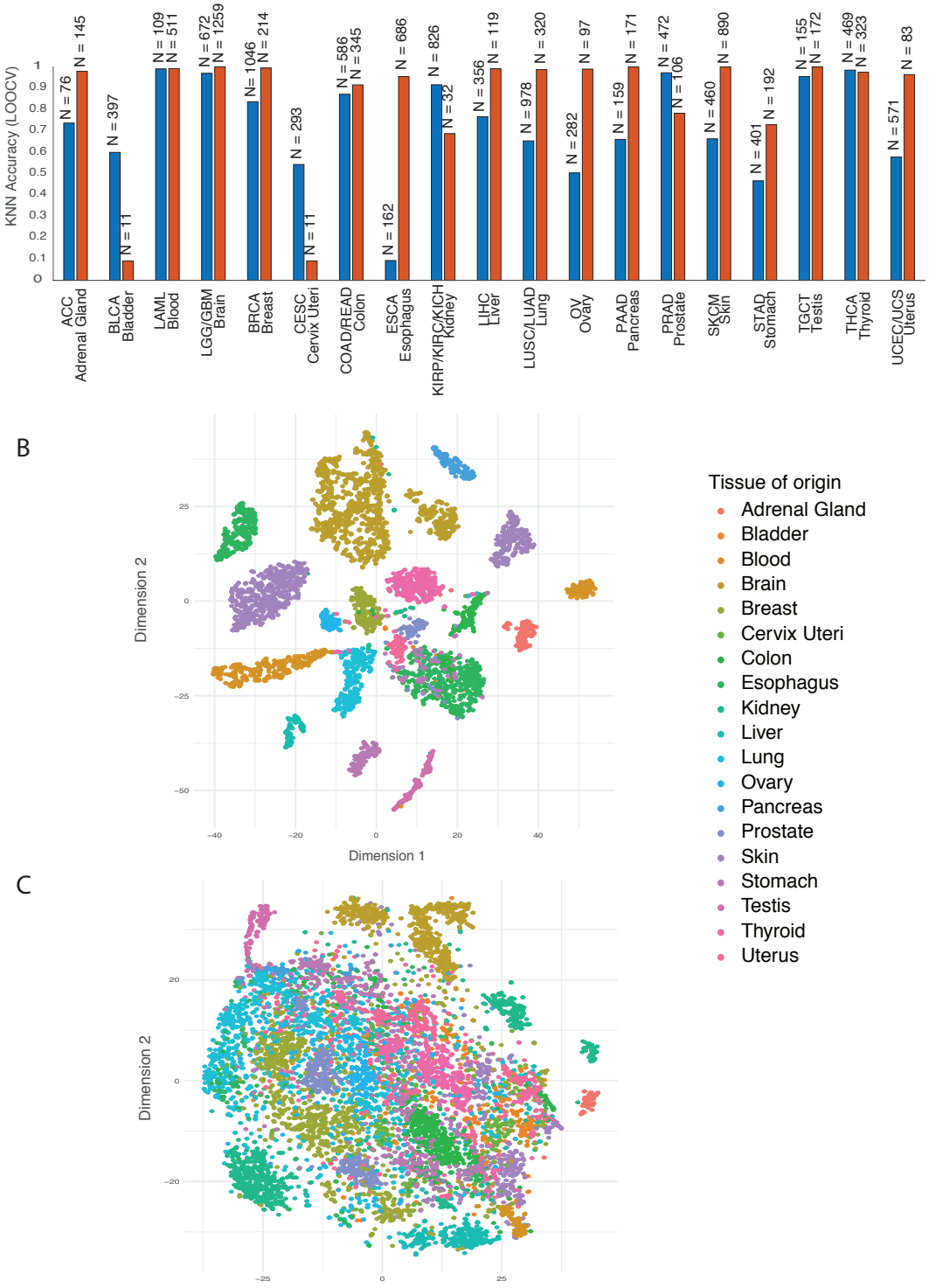

Figure 4.2: (A) K-Nearest-Neighbors (KNN) multi-class analysis: predictions made in a leave one out fashion (i.e., the accuracy). Height of bars indicate the fraction of correctly predicted cases. The numbers on top of each bar indicate the number of samples available for each class. (B,C) t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction analysis of chromosome arm-wide mean gene expression levels in normal tissues (B) and in cancers (C).

4.3.2 Chromosome arm imbalance scores of cancer types and the distribution of cancer type-specific driver genes over chromosome arms

Recent studies have looked at the connection between specific chromosomal gains and losses and driver genes located on these chromosomes for specific cancer types [55, 243]. In this study, we revisit this connection. For each tissue analyzed in this study, the correlation between the frequency of losses in associated cancer types and the fraction of drivers that are tumor suppressors is consistently strong and positive (Figure 4.3 panel A, permutation test with 1000 random shuffling of arms and tissue pairing of the values in Figure 4.3 panel A, p-value < 0.05) The strongest of these correlations are observed for chromosome arms 17p, 17q and 9p. The direction of correlation between gains of chromosome arms and the location of tissue specific oncogenes is however less clear (Figure 4.3 panel B, empirical p-value after 1000 iterations of random shuffling is > 0.05, Supplementary Table 5). To explore this further, we performed four hierarchical clustering analyses of tissues based on i) chromosomal arm imbalance scores in associated cancer-types (Figure 4.4 panel A), (ii) mean chromosome arm-wide gene expression levels in associated cancer types (Figure 4.4 panel B), (iii) mean chromosome arm-wide gene expression levels in normal tissue (Figure 4.4 panel C), and (iv) chromosome arm-wide imbalance in the fraction of oncogenes and tumor suppressor genes originating from each tissue (Figure 4.4 panel D). For ease of visualization, the tissues are partitioned

and coloured by 4 distinct clusters obtained from each hierarchical clustering separately. We find that the hierarchical clustering of tissues based on chromosomal arm imbalance scores (Figure 4.4 panel A) and the hierarchical clustering based on mean chromosome arm-wide normal gene expression levels (Figure 4.4 panel C) are highly similar (spearman correlation of cophenetic distances = 0.61, p-value < 2.2E-16). Likewise, a strong similarity is observed between hierarchical clustering of tissues based on arm imbalance scores (Figure 4.4 panel A) and hierarchical clustering based on cancer gene expression levels (Figure 4.4 panel B) (spearman correlation of cophenetic distances = 0.52, p-value = 1.57E-13). However, such a similarity is not observed when looking at the arm imbalance scores (Figure 4.4 panel A) and distribution of tissue specific oncogenes and tumor suppressor genes across arms (Figure 4.4 panel D) is (spearman correlation of cophenetic distances = -0.09, p-value = 0.2067). While the list of tissue specific cancer driver genes is still incomplete, these results suggest that copy number changes in resident driver genes may not be sufficient to explain the observed tissue-specificity of chromosomal aneuploidies in cancers.

4.3.3 Chromosome arm-wide methylation levels in normal tissues

A possible mechanism regulating chromosome-wide gene expression levels in normal tissues is DNA methylation. Therefore, in a fashion similar to Figure 1B, we explored whether mean chromosome arm-wide methylation levels correlate with the mean chromosome arm-wide gene expression levels. The Gene Expression Omnibus

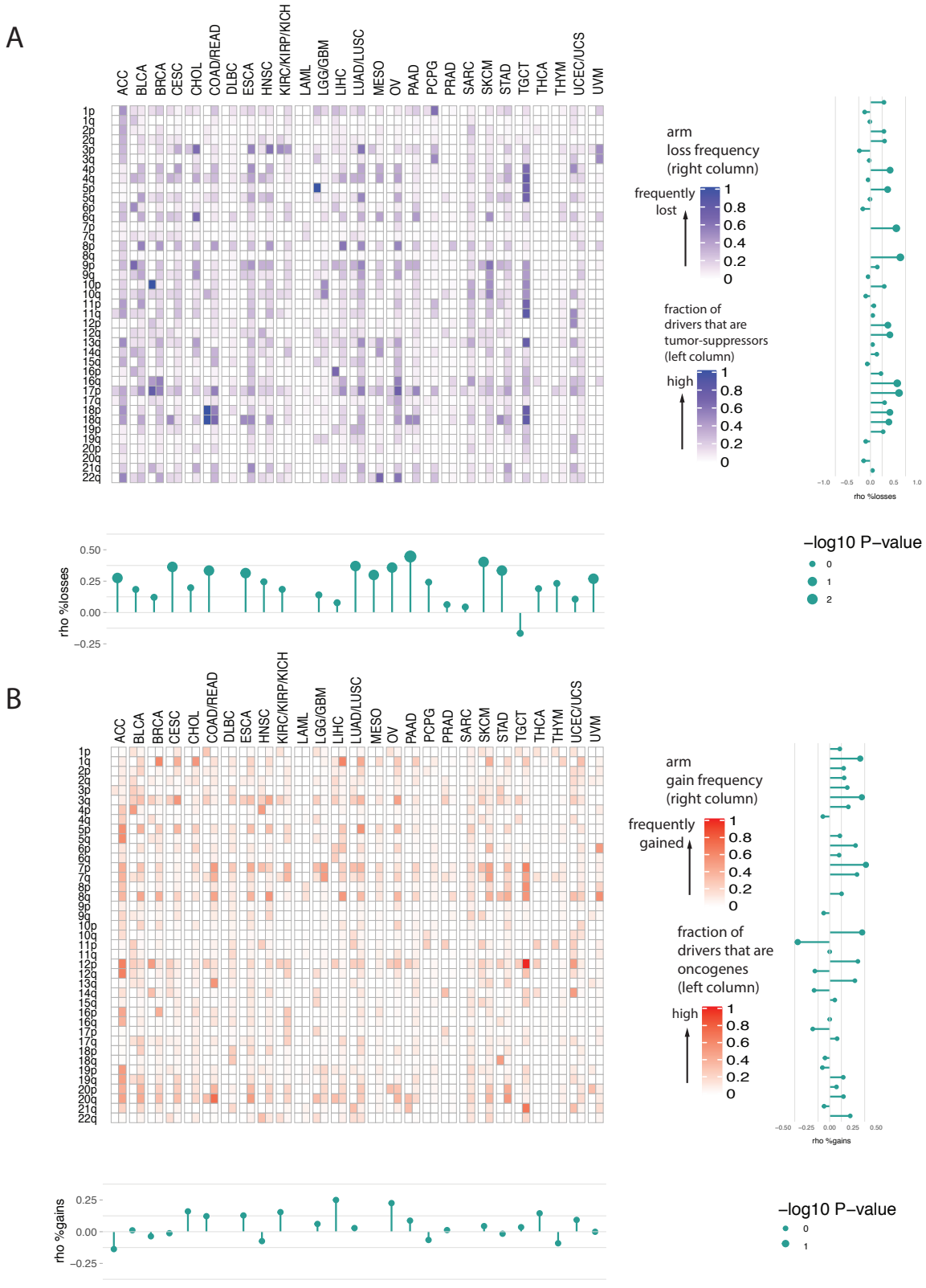

Figure 4.3: (Continued on the following page.)

Figure 4.3: For each set of cancer types with shared tissue of origin, we plot: A) The fraction of driver genes on each arm that are considered to be tumor suppressors (left column) and the frequency of losses reported for the arm. The bluer the color, the higher the tumor suppressor burden (and likewise for the frequency of losses). B) The fraction of driver genes on each arm that are considered to be oncogenes (left column) and the frequency of gains reported for the arm (right column). The redder the color, the higher the oncogenic burden (and likewise for frequency of gains). Barplots shown beside each heatmap are the spearman rank correlations (horizontal bars indicate coomparisons for each arm independently, vertical bars indicate comparisons for each tissue independently). The size of bubbles indicates the p-value. A size of 2 indicates p-value < 0.01, a size of 1 indicates p-value < 0.1 and size of 0 indicates p-values < 1. As seen at the tissue level, correlation between tumor suppressor burden and frequency of losses is almost always positive (permutation test p-value after randomly shuffling data < 0.05), whereas that is not the case for gains.

(GEO) database provides genome-wide methylation levels for 11 different tissue types, all obtained using the same Illumina 450K platform. Based on these data, we analyzed chromosome arm-wide mean methylation patterns for 11 tissues from 765 samples (Materials and Methods, Supplementary Table 5). For each tissue, we observe that differences in mean methylation levels across chromosomal arms within a tissue are consistently negatively correlated with corresponding mean arm-wide gene expression levels (permutation test with 1000 random shuffling of arms and tissue pairing of the values, p-value < 0.05) (Figure 4.5 panel A). However, for a single arm across tissues, the directionality of correlations are less consistent. This could potentially be due to the small number of tissues analysed. Furthermore, an individual sample-level classification analysis using the KNN algorithm reveals that one can predict (in leave one out cross-validation) the normal tissue of origin of individual samples just based on chromosome arm-wide mean methylation levels. The clustering of samples by tissue is visualized using t-SNE dimensionality reduction. (Figure 4.5 panels B and C). Tissues with very few samples had poor classification accuracy as expected from KNN. These results suggest that normal chromosome

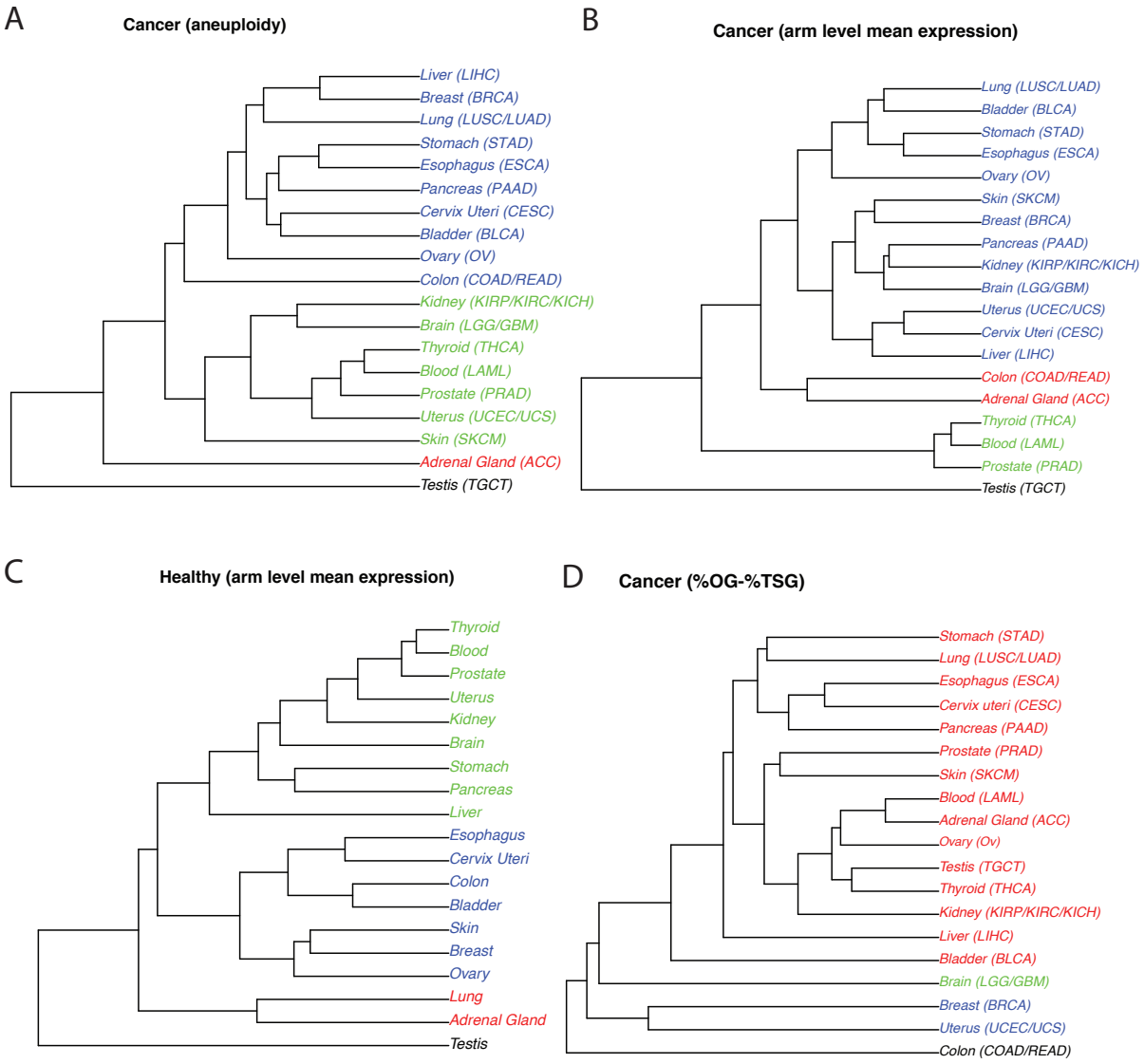

Figure 4.4: (A) cancer chromosome arm-wide gains and losses, (B) cancer mean chromosome arm-wide gene expression, (C) mean chromosome arm-wide gene expression of normal tissues, and (D) chromosome arm-wide imbalance of tumor suppressor genes and oncogenes. Note that the clusters are similar in A-C, yet different in D.

arm-wide methylation levels may play some part in regulating the transcriptional output of each chromosome arm.

4.4 Discussion

Chromosomal aneuploidies are a defining feature of tumors of epithelial origin. These aneuploidies result in tumor type specific genomic imbalances [116, 93, 187, 21, 22]. As of yet, there is no sufficient explanation for this specificity [21]. In this work, we systematically compared the frequencies of chromosome arm gains and losses in different cancer types to the mean chromosome arm wide gene expression leves in normal tissues of origin and distribution of known or implicated tissue-specific oncogenes/tumor suppressors across chromosome arms. Our analysis reveals a complex picture of factors driving frequent chromosome arm alterations in specific cancer types. Specifically, we notice recurrent losses in chromosome arms in cancer types where tissue-specific tumor suppressors reside, suggesting that these losses broadly target these driver genes. However, the targets of recurrent tissue-specific chromosomal gains are less clear. While it is possible that these chromosomal gains are targeting yet unidentified oncogenes, our analysis of normal chromosome wide gene expression and methylation data suggests an alternative paradigm in which these alterations instead aim to hardwire expected gene expression levels of normal tissue origin. This notion is further supported by recent observations across multiple cancer types where oncogenes were found to be preferentially activated via extra-chromosomal DNA [114]. The functional implications of many genes that are

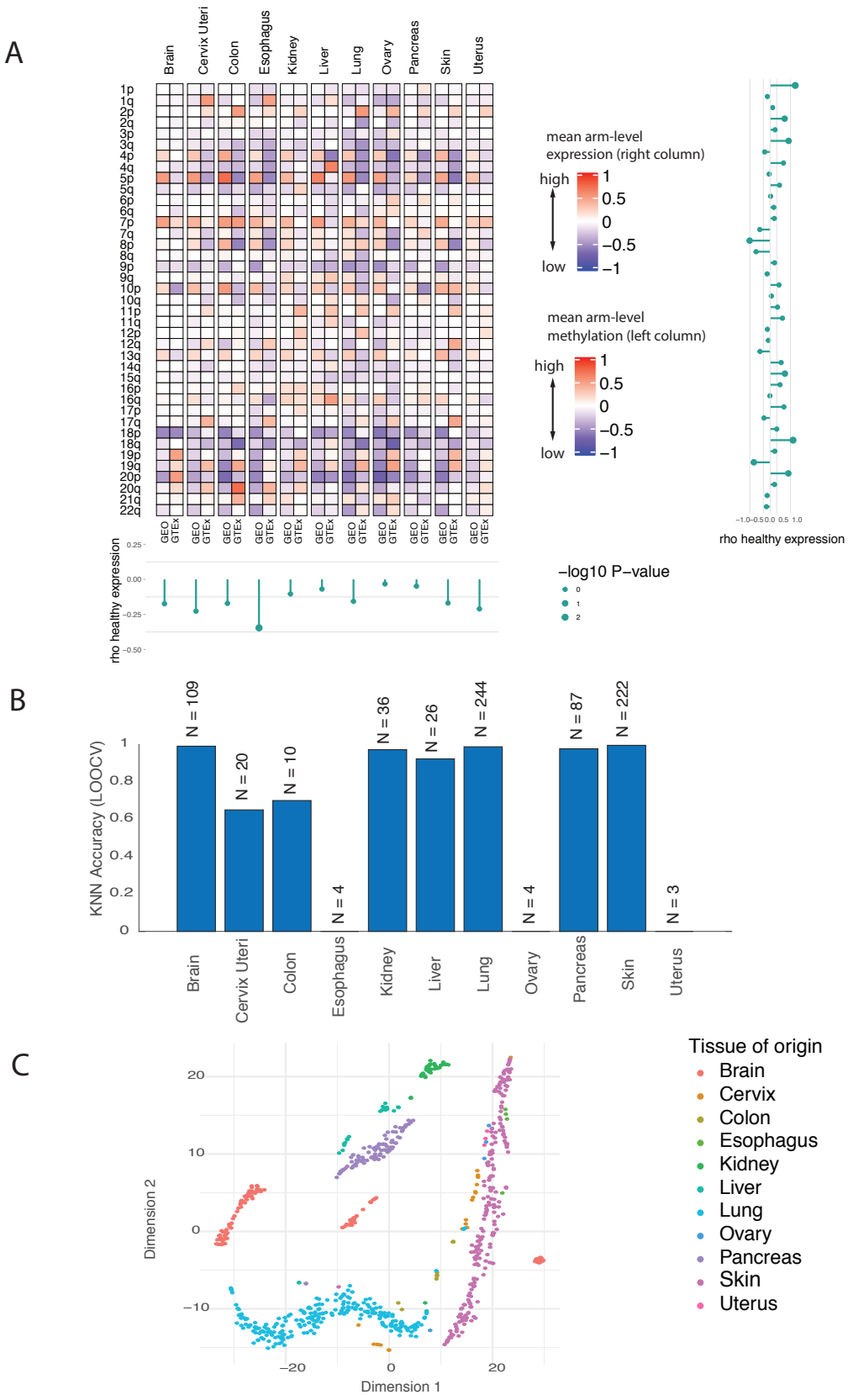

Figure 4.5: (Continued on next page.

Figure 4.5: (A) For each tissue with available normal methylation data, we plot the mean arm-wide methylation levels of each arm (left column) and the mean arm-wide expression levels of each arm (right column). The mean expression and methylation values are quantile normalized to the same scale (See Methods) for comparison and visualization. For left column: The redder the color, the higher the arm-wide methylation level, the bluer the color the lower, the lower the arm-wide methylation level . For right column: the redder the color, the higher the arm-wide expression level, the bluer the color the lower the arm-wide expression levels. Bar plots besides the heatmap are spearman rank correlations (horizontal bars indicate comparison for each arm independently, vertical bars indicate comparison for each tissue independently). The size of bubbles indicates the p-value. A size of 2 indicates p-value < 0.01, a size of 1 indicates p-value < 0.1 and size of 0 indicates p-values < 1. As seen at the tissue level, correlation between arm-wide methylation levels and expression levels is consistently negative (permutation test p-value after random shuffling the data < 0.05). (B) Leave One Out Cross-Validation Accuracy of predicting each tissue entity based on chromosome wide mean methylation levels of each sample. The height of the bar indicates the accuracy quantified as fraction of samples correctly classified. The numbers on top of each bar indicate the number of samples from a given tissue. (C) tSNE plot depicting the clustering of different tissue samples by chromosome arm wide mean methylation levels.

affected by these alterations remain incompletely understood. We previously showed experimentally that the gain of chromosome 13 in colorectal cancer activates both Notch and Wnt signaling [36], and that the acquisition of extra copies of chromosome 7 in normal colon cells results in upregulation of cancer-associated pathways [29], which could imply that tissue-type specific chromosome arm-wide gene expression levels promote cellular fitness. Of note, Sack et al. [200] have demonstrated that the inclusion of tissue-specific growth promoting genes strengthens the correlation between chromosome arm loss/gain ratios and the proliferation-driving capability of each chromosome-arm in breast and pancreatic cancers. Graham and colleagues reported a general role of copy number alterations and metabolic selection pressure [81]. Despite the ubiquitous presence of chromosomal aneuploidies in most solid tumors, there are also several publications pointing to a reduction of cellular fitness as a consequence of general aneuploidy in model systems such as yeast, immortalized murine embryonic fibroblasts and typically near-diploid cancer cells engineered to

harbor specific trisomies [82, 244, 209], so the functional implications of these events remains an open challenging question.

There are some limitations specific to the data analysis conducted in this study. Firstly, our analyses comparing cancer types to normal tissues were restricted to tissues where data was measured in a homogeneous fashion on the same platform and publicly available (i.e., GTEx for gene expression and GEO for methylation). Furthermore we restricted ourselves to external data sources for normal tissue expression and methylation rather than use adjacent normal tissue samples from the TCGA. This was mainly due to incomplete availability of methylation and expression of normal adjacent to tumor samples for many cancer types and the presence of stromal and immune cell contamination in these tissues [9, 98]. Secondly, identification of existing and potentially new cancer type specific oncogenes and tumor suppressors was previously done by combining evidence from multi-omic sources into one prediction score using supervised machine learning [15]. However, this list is still incomplete and the mechanism of action of many of these genes in different cancer types is not completely understood. Thirdly, since we are exploring correlation patterns across different tissue and cancer types, it is likely that more significant associations would be observed in arms with specific, high-intensity trends of either gain or loss compared to arms that are less frequently altered. In sum, our data analysis suggests that chromosome aneuploidies could be potentially involved in the maintenance of gene expression levels characteristic of the normal tissue of origin of cancers, in addition to targeting cancer type specific driver genes (Figure 4.6).

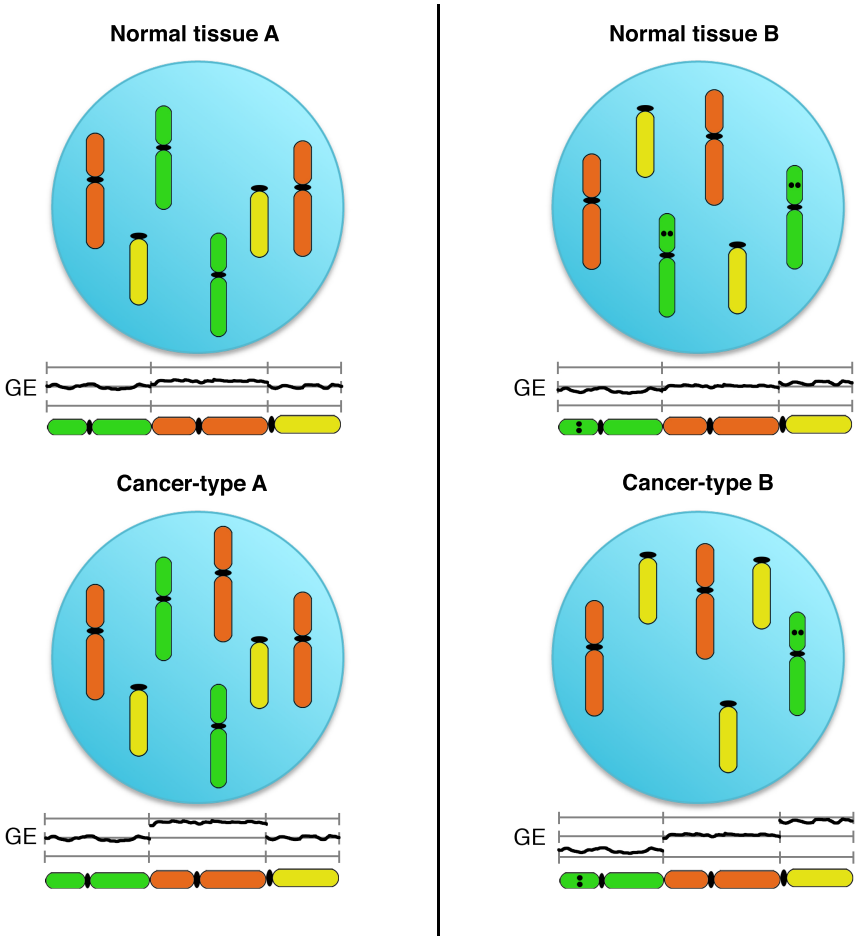

Figure 4.6: Genes on the red chromosomes are on average, expressed at slightly higher levels compared to other chromosomes in normal tissue A, whereas in normal tissue B, the yellow chromosomes shows increased tissue-specific expression on average and genes on the green chromosome are expressed at lower levels on average. The acquisition of chromosomal aneuploidies in the respective cancer-types (gain of the red chromosome in cancer-type A and the yellow chromosome in cancer-type B, accompanied by the loss of the green chromosome in cancer-type B amplifies this effect and provides the genetic basis of "hard-wiring" tissue-specific chromosome arm-wide gene expression levels. The dots on the green chromosome reflects the presence of a tumor suppressor gene, which can be targeted by the loss of the green chromosome in tumor evolution.

Chapter 5

Algorithms for dissecting cellular heterogeneity in the TME (Tumor Micro Environment)

⋆⋆ This work was done in collaboration with Dr. Kun Wang. A bioarxiv pre-print is available [240].

5.1 Overview

The importance of the tumor microenvironment (TME) in cancer has been recognized since the late 1800s [169]. The recent success of immune checkpoint blockade has further sparked interest in studying TME interactions that shape clinical outcomes following immunotherapy, aiming to find biomarkers of treatment response and new treatment opportunities [216]. One key step in studying these interactions is the characterization of the molecular profiles of different cell types in a patient's tumor sample. Fluorescence-activated cell sorting (FACS) and single-cell RNA sequencing have emerged as effective tools to address this challenge [238]. However, due to the cost of these procedures and scarcity of fresh tumor biopsies, the application of these approaches has remained limited. Given that bulk tumor gene expression from preserved biopsies is far more abundant, computational methods that can effectively extract cell-type-specific expression from such data, termed deconvolution algorithms, could be very helpful. If successful, such deconvolution

methods can markedly advance our knowledge of the TME across many tumor types and different contexts, and beyond that, they may be readily applied to interrogate other large bulk expression datasets.

Several previous studies have developed a variety of expression deconvolution algorithms. DeMixT [3] was designed to estimate individual-specific expression for three cell components provided prior reference samples of two of these cell components. ISOpure [181] has aimed to derive sample-specific cancer cell expression with the assumption that the observed bulk gene expression profile is a mixture of predefined stromal and immune cell expression profiles that are shared across all the samples. Building on this work, Fox et al extended ISOpure to predict individual-specific non-tumor cell expression by subtracting cancer cell profiles from the bulk mixtures in a two-cell type model [65]. More recently, Newman et al[163] developed CIBERSORTx, the first approach that aims to predict the sample-specific gene expression of all cell types composing it by employing a set of novel deconvolution heuristics. As a proof of concept, Newman et al showed that CIBERSORTx can accurately reconstruct the cell-type-specific expression of genes in each input sample under certain modelling assumptions. This groundbreaking work has, however, some notable limitations: (1) The number of genes whose cell-type-specific expression can be reconstructed in each sample is relatively small, especially for low-abundance cell types, and (2) their approach does not provide confidence estimations of the predictions made, while such estimations could be potentially useful in most deconvolution applications in the absence of ground truth data.

Here, we introduce a new deconvolution algorithm and software, CODEFACS

(COnfident DEconvolution For All Cell Subsets), which markedly advances the ability to successfully deconvolve bulk gene expression data. CODEFACS receives as input bulk gene expression profiles of tumor samples and either pre-computed estimates of abundance of expected tumor, immunological and stromal cell types in each sample, or their prototypical molecular signatures, which serve as seeds for estimating the abundance of each cell type in each sample. CODEFACS then predicts the cell type-specific gene expression profiles of each sample. It is a heuristic approach aimed at maximizing the number of genes in each cell type whose expression across the samples can be confidently predicted via a heuristic method to estimate confidence. Using 15 benchmark datasets where the ground-truth is known, we show that CODEFACS robustly improves over CIBERSORTx, both in terms of gene coverage and the individual gene expression estimation accuracy. We additionally developed LIRICS (LIgand Receptor Interactions between Cell Subsets), a pipeline that integrates the output of CODEFACS with a database of prior immunological knowledge that we curated to infer the active cell-cell interaction landscape in each sample. These data can then be analyzed in conjunction with any sample-associated clinical annotations (e.g., response to treatment) to infer the most important clinically relevant immune interactions between the cell types in a given patient's cancer cohort.

Building on its enhanced coverage and accuracy, we next applied CODEFACS to reconstruct the cell-type-specific transcriptomes of 8000 tumor samples from 21 cancer types in TCGA. Analyzing these fully deconvolved TCGA expression datasets using LIRICS we find a shared repertoire of intercellular interactions enriched in the

TME of mismatch repair deficient tumors of different tissues of origin, which is associated with improved overall patient survival and high response rates to anti-PD1 treatment, independently of their mutation burden levels. Finally, using machine learning techniques, we identify a subset of intercellular TME interactions that are predictive of response to immune checkpoint blockade treatment in melanoma patients.

In summary, CODEFACS and LIRICS present a new way to analyze large bulk RNA-seq datasets to study cellular crosstalk in the TME of each patient, and to learn more about the association of different tumor-immune interactions with different clinical measures. The potential scope of applications of both CODEFACS and LIRICS goes beyond studying the TME, as these tools can be applied to study any disease of interest given bulk gene expression data and relevant reference signatures of cell types involved.

5.2 Methods

5.2.1 Data curation

5.2.1.1 Single cell RNA-seq datasets

To benchmark the performance of CODEFACS, we first set out to obtain publicly available single cell RNA-seq datasets where both tumor and non-tumor cells were successfully isolated. This search led us to the identification of nine such single cell RNA-seq datasets from the literature, each from a different cancer type.

Collection of additional single cell datasets was frozen after Dec 2019. For each dataset sequenced on the SmartSeq2 platform, the log normalized transcript counts for each gene in each sequenced cell were made publicly available by the original authors. For the application of deconvolution, these counts were transformed back to the Transcripts Per Million (TPM) scale. For datasets sequenced on the 10x platform, UMI counts for each gene were made publicly available and were scaled by the library size of each cell and multiplied by a factor of 1 million to get expression values in TPM scale. See Table 5.1

Table 5.1: Single cell RNASeq datsets collected and analyzed in this study

Dataset Reference	**Cancer type**	**Sequencing Platform**	**Number of patients studied**
GSE115978	Melanoma	SmartSeq2	32
GSE131928	GBM	SmartSeq2	28
GSE103322	HNSCC	SmartSeq2	18
CRA001160	PDAC	10x	35
GSE125449	LIHC	SmartSeq2	12
GSE81861	CRC	SmartSeq2	11
E-MTAB-6149	LUAD	10x	3
E-MTAB-6149	LUSC	10x	2
GSE118389	TNBC	SmartSeq2	6

5.2.1.2 Bulk RNA-seq datasets

Gene expression and matching bulk tumor methylation data from fresh frozen tumor biopsies in TCGA were downloaded from [77]. In addition, publicly available bulk expression data from formalin fixed paraffin embedded tumor biopsies of melanoma patients receiving immune checkpoint blockade treatment were downloaded from [185, 75, 131]. All bulk RNA-seq datasets were collected such that they

have a sufficiently large sample size to reliably perform complete deconvolution of expression profiles (¿ 4 times the number of cell-types involved) [163]. Collection of datasets was frozen after Dec 2019. To maintain consistency with the pipeline used for preprocessing TCGA data, bulk gene expression levels in immune checkpoint blockade datasets were re-quantified using STAR v2.7.6a and RSEM v1.3.3 [30] with GENCODE v23 human genome annotation [91]. Furthermore, to mitigate technical biases, between-sample scaling factors were estimated using TMM method implemented in edgeR [195] and TPM values in each sample were further rescaled by these scaling factors [196].

Table 5.2: bulk RNASeq datasets collected and analyzed in this study

Cohort Name	Cohort description
The Cancer Genome Atlas	6972 samples spanning 21 distinct cancer types in the TCGA with matched bulk methylation profiles. All samples were biopsied pre-treatment.
Riaz et al, Cell, 2017 [185]	109 samples from 73 patients, 51 patients had pre-treatment samples (25 anti- PD1 monotherapy, 26 previously progressed on anti-CTLA4)
Gide et al, Cancer Cell 2019 [75]	91 samples from 75 patients, 73 patients had pre-treatment samples (41 anti-PD1 monotherapy, 32 anti-PD1+anti CTLA4)
Liu et al, Nature Medicine 2019 [131]	121 samples from 121 patients, 120 patients had pre-treatment samples (75 anti-PD1 monotherapy, 45 previously progressed on anti-CTLA4)

5.2.1.3 Generation of simulated bulk RNA-seq datasets

To evaluate the performance of CODEFACS, we generated 14 different pseudo-bulk RNA-seq datasets from mixing experiments with single cell data. Each sample in each benchmark dataset has matching cell type specific gene expression profiles derived from averaging single cell RNA-seq profiles of individual cells from the same sample and same cell type. These profiles serve as the ground truth for the eval-

uation of deconvolution performance. To avoid any circularity in our validations, for each of the single cell datasets involved, single cell data from 4 randomly chosen patients were separated from the rest. These data were used to derive reference gene expression signatures for each cell type. The mixing experiments were then performed on single cell data of the remaining patients that were hidden from the reference signature derivation process. In addition, we simulated technical replicates for each pseudo-bulk sample, wherein we injected noise in the pseudo-bulk expression of a few randomly chosen genes and then renormalized the expression data by the sample library size. This procedure simulates mRNA composition noise that is commonly observed in bulk RNA-seq datasets due to technical differences in sample preparation [203, 64, 227, 194]. In addition, we obtained a FACS sorted lung cancer dataset which include purified RNA-seq for four cell types 10 and generated a pseudo bulk correspondingly [72]. In total, 15 benchmark datasets were generated.

Table 5.3: List of 14 artificially generated bulk expression datasets with matched cell-type specific expression measurements for each sample (Used for performance evaluation of CODEFACS and CIBERSORTx).

Benchmark Dataset name	**Description**
SKCM dataset 1	28 pseudo bulk melanoma samples generated by averaging single cell RNASeq expression profiles (GSE115978) from the same patient
SKCM dataset 2	28 pseudo bulk melanoma samples generated by averaging imputed single cell RNASeq expression profiles (GSE115978) from the same patient. Imputation of single cell RNASeq data was performed using scImpute v0.0.9
SKCM dataset 3	First noisy technical replicate of SKCM dataset 2 generated by injecting noise in each of the 28 mixes
SKCM dataset 4	Second noisy technical replicate of SKCM dataset 2 generated by injecting noise in each of the 28 mixes
SKCM dataset 5	100 pseudo bulk melanoma samples generated by sampling single cell RNASeq expression profiles (GSE115978) from different patients in varying cell type specific proportions and averaging them.
SKCM dataset 6	First noisy technical replicate of SKCM dataset 5 generated by injecting noise in each of the 100 mixes
SKCM dataset 7	Second noisy technical replicate of SKCM dataset 5 generated by injecting noise in each of the 100 mixes
GBM dataset 1	24 pseudo bulk GBM samples generated by averaging single cell RNASeq expression profiles (GSE131928) from the same patient
GBM dataset 2	24 pseudo bulk GBM samples generated by averaging imputed single cell RNASeq expression profiles (GSE131928) from the same patient. Imputation of single cell RNASeq data was performed using scImpute v0.0.9
GBM dataset 3	First noisy technical replicate of GBM dataset 2 generated by injecting noise in each of the 24 mixes
GBM dataset 4	Second noisy technical replicate of GBM dataset 2 generated by injecting noise in each of the 24 mixes
GBM dataset 5	100 pseudo bulk GBM samples generated by sampling single cell RNASeq expression profiles (GSE131928) from different patients in varying cell type specific proportions and averaging them.
GBM dataset 6	First noisy technical replicate of GBM dataset 5 generated by injecting noise in each of the 100 mixes
GBM dataset 7	Second noisy technical replicate of GBM dataset 5 generated by injecting noise in each of the 100 mixes

5.2.1.4 Curation of reference signatures of cell types

For the application of CODEFACS, molecular profiles of signature genes of each cell type of interest are needed to estimate the relative cell fractions in the bulk. We used single cell expression derived signatures as priors to deconvolve the melanoma ICB datasets. To derive these signatures from single cell data, we first start out by obtaining the class labels of each cell type of interest. These data are publicly available for each single cell dataset we collected. Hence, we primarily use these labels in our study (unless further refinement of labels into specific cell subtypes of interest is needed for a specific usage). With a collection of single cell expression profiles and matching cell type labels as input, we used CIBERSORT online tool to derive a cell-type-specific signature matrix. Thereafter, we applied CODEFACS to ICB datasets with default parameters settings and batch correction requirement specified. For TCGA deconvolution, we first estimated cell fractions based on bulk methylation and then applied CODEFACS to corresponding bulk gene expression for the 21 cancer types which have both types of data available. We chose methylation signatures over expression-based signatures for TCGA analysis for two reasons. First, single cell expression data with consistent cell types across 21 cancer types are not available. Second, DNA methylation-based signatures are considered to be more stable marks of cellular identity compared to dynamic RNA expression derived signatures [24]. The methylation-based cell type signatures were obtained from MethylCIBERSORT [39]. We applied CODEFACS to TCGA datasets with default parameters settings and without batch correction requirement specified.

Table 5.4: list of all cell types with reference methylation signatures available from MethylCIBERSORT

non-cancer cell types with methylation signatures	
Endothelial	
Fibrolast	
CD14+ (Monocyte/Macrophages/Dendritic cells)	
CD19+ (B cells)	
CD56+ (NK cells)	
CD4+ (T cells)	
CD8+ (T cells)	
Treg (T cells)	
Eos (Eosinophils)	
Neu (Neutrophils)	
cancer cell types with methylation signatures	**Matched cancer type (TCGA)**
endometrium	TCGA-UCEC
large_intestine	TCGA-COAD
stomach	TCGA-STAD
mesothelioma	TCGA-MESO
breast	TCGA-BRCA
oesophagus	TCGA-ESCA
kidney	TCGA-KIRC
sarcoma	TCGA-SARC
head_and_neck	TCGA-HNSC
prostate	TCGA-PRAD
liver	TCGA-LIHC
lung_NSCLC_adenocarcinoma	TCGA-LUAD
lung_NSCLC_squamous_cell_carcinoma	TCGA-LUSC
bladder	TCGA-BLCA
skin	TCGA-SKCM
glioma	TCGA-GBM
pancreas	TCGA-PAAD
thyroid	TCGA-THCA
acute_myeloid_leukaemia	TCGA-AML
B_cell_lymphoma	TCGA-DLBC

5.2.2 Full in-silico deconvolution of bulk mixtures

CODEFACS is designed to do the following:

Input. *(i) Bulk gene expression of a collection of samples (required) (ii) Cell fraction*

estimates of expected tumor, immune and stromal cell types in each sample; OR Cell-type-specific signature profile (required if cell fractions are not provided)

Goal. *(i) Predict the expression of each gene in each sample in each cell type in the mixture (ii) Estimate confidence scores [0-1] for each gene-cell-type pair, which denote the confidence level in the predicted expression of a gene in a cell type across samples (≈1 High confidence, ≈0 Low confidence)*

In this section, we provide a formal description of the computational problem being solved by CODEFACS. The full deconvolution problem is formulated as follows:

$$\begin{aligned} \min \quad & \textstyle\sum_{i=1}^{m} \left\| \left(B_{i,\cdot} - \mathrm{diag}(G_{i,:,\cdot} \times F^T) \right) \right\| \\ \text{s. t.} \quad & \textstyle\sum_{k=1}^{c} f_{jk} = 1 && \forall j \\ & g_{ijk} \geq 0 && \forall i, j, k \\ & f_{jk} \geq 0 && \forall j, k \end{aligned} \tag{1}$$

where B represents the given bulk RNA-seq expression matrix (m genes $\times$ n samples), in which each entry b_{ij} is the observed bulk expression for i^{th} gene and j^{th} sample; G is a three-dimensional deconvolved gene expression matrix (m genes $\times$ n samples $\times$ c cell types), in which g_{ijk} denotes the unknown expression for gene i in the j^{th} sample and k^{th} cell type; F is the cell fraction matrix (n samples $\times$ c cell types), in which f_{jk} denotes the unknown cell fraction of k^{th} cell type in j^{th} sample. F varies across samples and cell types but is constant across genes. $\|.\|$, represents the

L2-norm (which measures the reconstruction error) and diag() represents a function that gives a vector by extracting the diagonal entries of a matrix. The objective is to find an optimal solution for G and F with the constraint that the cell fractions (of c cell types) in any sample j sum up to 1 and all the gene expression values g_{ijk} are non-negative real values. In this study, we assume the gene expression is quantified as TMM normalized TPM values. More specially, we employed a strategy introduced by Monaco et al. [155], which first estimates between-sample scaling factors upon raw TPM values using TMM method [196] and further scale TPM values in each sample using these scaling factors.

Problem (1) has no unique optimal solution without additional constraints and regularizations since there are more parameters to be estimated than observations [181, 163]. However, problem (1) can be separated into two independent problems: cell fraction estimation and cell-type-specific gene expression prediction for each individual sample. The cell fraction estimation problem is formulated as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^{l} \left\| B'_{i,\cdot} - \left(S_{i,\cdot} \times F^T \right) \right\| \\ \text{s. t.} \quad & \sum_{k=1}^{c} f_{jk} = 1 \qquad \forall j \\ & f_{jk} \geq 0 \qquad \forall j, k \end{aligned} \tag{2}$$

Where S denotes the cell-type-specific signature matrix (l genes x c cell types) and the l genes are a subset of all the m genes in G or B matrix that are preferentially over-expressed in at least one of the c cell types and their expression is assumed to be constant across the population to arrive at an approximate solu-

tion of cell fractions in each sample. F is the same as that in equation (1), while B' is a submatrix of the bulk expression matrix B in equation (1) corresponding to the l genes in cell-type-specific signature matrix S. Numerous effective cell fraction estimation tools have been developed and reported to solve problem (2) [211, 69, 179, 257, 79, 2, 162, 241, 166, 129, 154, 16, 175, 61, 104, 8, 255]. The experimental analog to these methods is the cell gating procedure described in FACS. We solve this problem using a well-known reference-based approach: CIBERSORT [162]. If needed, CODEFACS also provides a batch correction approach introduced by CIBERSORTx that could be applied to minimize cross-platform technical batch effects between bulk mixture profile and cell type signature profile generated from different technical platforms (e.g. bulk RNA sequencing, SmartSeq2-based single cell sequencing, 10x-based single cell sequencing and microarray expression profiling) [163]. In addition, we provide the option to input prior known cell fractions instead of performing cell-fraction estimation de novo. Either known cell fractions or cell type signature profiles are required as input. Newman et al. [162, 163] found that cell fractions determined by the CIBERSORT algorithm, which we reimplement in CODEFACS, mostly exhibit strong concordance with ground truth.

Once F is estimated or provided, the full deconvolution problem formulated in (1) can be reduced to solving for G, given B and F. One can additionally reduce problem (1) to a simpler problem where one solves for the expected cell-type-specific expression for a specific gene across a group of individual samples, given the cell fractions and bulk expression matrix:

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} \left\| B_{i,.} - \left(\bar{E}_{i,.} \times F^T \right) \right\| \\ \text{s. t.} \quad & \bar{e}_{ik} = 1 \ \forall i, k \end{aligned} \tag{3}$$

where B is the same as in equation (1) and represents the input bulk expression matrix; F is also the same as that in equation (1) and denotes cell fractions; $\bar{E}$ is the expected cell-type-specific expression matrix (m genes $\times$ c cell types) across the population, in which $\bar{e}_{ik}$ denotes the expected expression of gene i in cell type k. For a fixed F, a unique optimal solution for this problem exists and can be found using non-negative least squares (NNLS) [163, 32, 168]. The key difference between problem (3) and problem (1) is that the former aims to predict expected cell-type-specific expression for each gene in the population, while the latter predicts the expected cell-type-specific expression for each gene in each sample.

One can aim to solve problem (1) approximately by making use of a greedy divide and conquer strategy that breaks down problem (1) into simpler problems (2) and (3). Newman et al, in their groundbreaking work CIBERSORTx, were the first to propose such an algorithm. In CODEFACS, we introduce the concept of confidence scores and additional algorithmic improvements to extend this approach. We show that CODEFACS yields a much more accurate solution compared to CIBERSORTx in 15 benchmark datasets with ground truth data.

The CODEFACS algorithm consists of three modules that are executed sequentially and a confidence ranking system that is invoked after the execution of

each module. In module 1 we refined and extended the high-resolution deconvolution module introduced by CIBERSORTx. First, we generalized their two-freedom estimation method into a recursive splitting method, which we call "p-freedom estimation" (The degrees of freedom represent the distinct latent sources of variability in gene expression across individuals). We found that p-freedom estimation could capture tumor heterogeneity better than the 2-freedom estimation. Second, we generalized their sliding window method by employing an ensemble of window sizes. Using an ensemble of window sizes seeks to reduce the dependence of downstream biological analyses on arbitrary choices of the window size parameter. In addition, we developed modules 2 and 3 (hierarchical deconvolution and imputation-based deconvolution) to further increase the number of highly predictable genes. The confidence ranking system uses a series of heuristics to decide where the solution can be improved by subsequent modules. See Figure 5.7 for the schematic diagram with the inputs and outputs.

5.2.3 The notion of confidence

Before we formally describe the algorithm, we introduce the concept of confidence, which is a central part of the algorithm. Each of the three prediction modules operates under specific modeling assumptions that are, in theory, uniformly applicable to all genes. However, in practice, certain genes might violate these assumptions. Therefore, for such genes, one cannot confidently say whether their predicted cell-type specific expression levels closely reflect the ground truth. To quantify this

uncertainty, we designed a confidence ranking system, which can decide whether a specific prediction requires further refinement in subsequent modules by defining a ranking Φ over genes for each cell type using confidence relevant features (more details are provided in following subsections). Additionally, the confidence ranking system also re-evaluates the confidence level of each final prediction (gene-cell type pair) and provides in the end report a confidence score between 0 and 1.

5.2.4 The CODEFACS Algorithm

5.2.4.1 Cell fraction estimation (optional)

Cell type signatures are derived based on prior reference datasets using the signature derivation module from CIBERSORTx. Thereafter, we implemented a support vector machine (SVM)-regression-based method to predict cell fraction given bulk expression/methylation and prior cell-type-signature profiles following the CIBERSORT algorithm21. Given the bulk mixture and cell type signature profile, the SVM regression model outputs predicted cell fraction for each cell type and sample (Figure 5.1). If the user provides prior known cell fractions as input, CODEFACS will skip this optional step.

5.2.4.2 Batch Correction to refine cell-fractions (optional)

To account for any systematic batch effects between bulk expression and independently generated cell-type-specific signature expression data, which could bias cell-fraction estimates, we re-implemented the batch correction method introduced

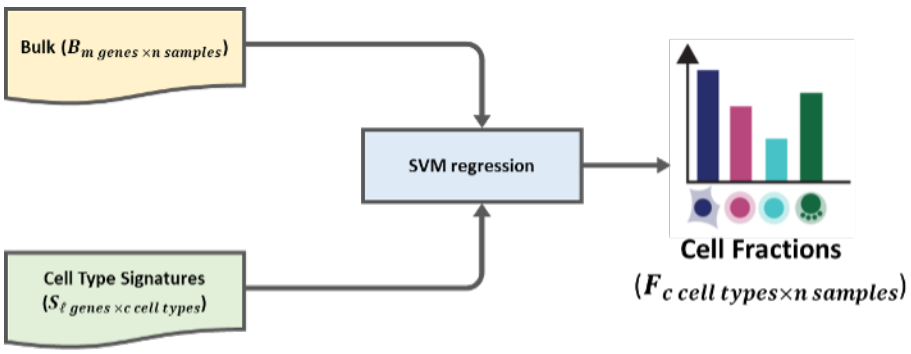

Figure 5.1: Given the bulk mixture matrix B (m genes $\times$ n samples) and cell type signature profile S (l genes $\times$ c cell types), we use SVM regression to predict the cell fraction matrix F (c cell types $\times$ n samples).

by CIBERSORTx [163]. The rationale for this method is that any batch effect between the given bulk expression and independently generated cell-type-specific signature expression must also be reflected in the reconstructed bulk expression $S \times F$. Thus, one can further refine cell-fraction estimates of each cell type in each sample after reducing the batch effect between the given bulk matrix and reconstructed bulk matrix $S \times F$, using the function ComBat() from the SVA package [125] in R (Figure 5.2). The final output of this step is a refined cell-fraction matrix. Currently our implementation focuses on correcting biases among bulk RNA sequencing and SmartSeq2-based single cell sequencing datasets. This step is optional and will be skipped if the user does not specify that it should be done. For more details on the batch correction procedure, please refer to section "Cross-platform normalization schemes for deconvolution" in the supplementary information of CIBERSORTx [163].

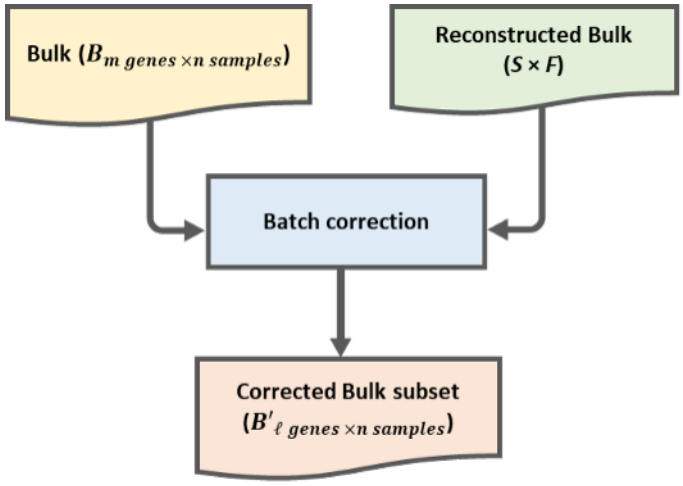

Figure 5.2: Given initial estimates of cell fractions F (c cell types $\times$ n samples) and cell type signature profile S (m genes $\times$ c cell types), one can reconstruct the bulk expression matrix via the matrix multiplication $S \times F$. Batch effects are then reduced between the given bulk subset B' and reconstructed bulk $S \times F$.

5.2.4.3 Module 1 - High resolution deconvolution

In this module, the observed bulk expression of a gene in a sample is modeled as the weighted sum of cell-type-specific expression of that gene from that sample (See problem 1 above).

Determine cell types in which a specific gene is weakly expressed (Step 1.1): To determine if a gene i is weakly expressed in a cell type, we first conduct the following statistical analysis: individuals are randomly chosen without replacement to generate 100 random subsets of individual samples and then problem (3) is solved to estimate expected cell-type-specific expression for each random subset. This bootstrapping procedure generates a distribution of expected cell-type-specific expression values $\bar{e}_{ik}$ in the population. We then derive two p-values for each cell type k: first, an empirical p-value that is estimated by checking the percentage of solutions where $\bar{e}_{ik} > 0$, and second, a p-value derived from a parametric t-test. The

two p-values are then combined using Fisher's method28 to obtain a final p-value for each cell type. If a gene is weakly expressed in a cell type (FDR > 0.2), we force the cell fractions of that cell type in the corresponding mixture model to be 0 to improve the deconvolution of gene expression in other cell types.

Recursively splitting samples into finite sub-groups (Step 1.2): With an appropriate cell-type-mixture model defined for each gene, we now try to find an approximate solution to problem (1). For a gene i, one can divide problem (1) into a finite number of simpler problems by assuming that individuals with similar bulk expression levels of gene i must have similar cell-type-specific expression levels of gene i. Hence, we first sort all samples in increasing order according to the bulk expression of gene i. In the two-freedom deconvolution in CIBERSORTx algorithm, one can then find a position t to partition all the sorted samples into two sorted subsets: $h_1 = 1, 2, \ldots, t-1$ and $h_2 = t, t+1, \ldots, n$, such that the expected cell-type-specific expression in each of the sub-sets (obtained from solving problem 3 using NNLS) best reconstructs the observed bulk expression (For more details, see section "Cell type expression coefficients that best explain the bulk GEP" in the supplementary information of CIBERSORTx [163]). Either of these two subsets can now be recursively partitioned further into smaller subsets in a similar fashion if the re-construction error keeps dropping and the subsets sample size stays above 1.9 times the number of cell types. This is referred to as p-freedom approach which extends the two-freedom approach of CIBERSORTx (Figure 5.3). The recursive splitting pseudo-code is shown below:

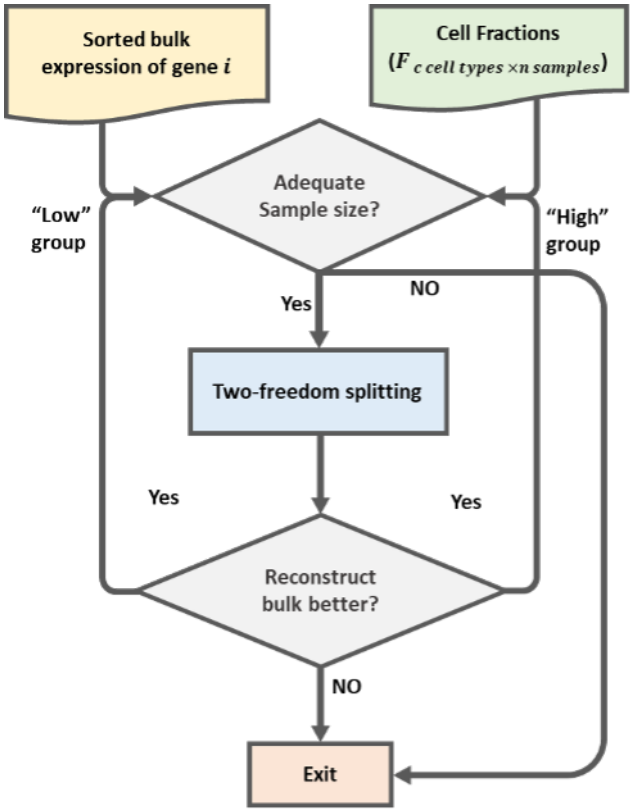

Figure 5.3: Given the estimated cell fractions F (c cell types $\times$ n samples) and sorted bulk expression of gene i, we check whether we have an adequate sample size for NNLS first: if not, it will exit; if yes, two-freedom splitting will be performed. Subsequently, we will check whether the two-freedom splitting improves the bulk reconstruction. If yes, both the low-expressed and high-expressed groups will recursively enter another round of two-freedom splitting; if no, the two-freedom splitting based predictions will be ignored and the function exits.

Algorithm 1 Recursive Splitting

Input:

F= cell fraction matrix
B = bulk expression matrix
size = the number of samples required for NNLS
i = gene index
st = start position in the sorted list of samples
ed = end position in the sorted list of samples
$\bar{E}_{i,.}$ = expected cell-type-specific expression of gene i over all sorted samples in the range st,. . . ,ed (Obtained from solving problem 3)

Output:

$G^p_{i,.,.}$ = cell-type-specific expression over samples $(1, 2, \ldots, n)$ from recursive splitting deconvolution

procedure RECURSIVE SPLITTING
  if $ed - st + 1 < 2 \times$ size **then**
    save $\bar{E}_{i,.}$ to $G^p_{i,.,.}$
    exit
  else
    $t \leftarrow$ 2-freedom index that splits sorted samples in the range $st, \ldots, ed$, into two subsets: $st, \ldots, st + t$ and $st + t + 1, \ldots, ed$ (See CIBERSORTx algorithm for more details)

    $\bar{L}_{i,.} \leftarrow$ expected cell-type-specific expression of gene i over all sorted samples in the range $st, \ldots, st + t$ (Obtained from solving problem 3)

    $\bar{H}_{i,.} \leftarrow$ expected cell-type-specific expression of gene i over all sorted samples in the range $st + t + 1, \ldots, ed$ (Obtained from solving problem 3)

    $\text{Err1} = \left\| B_{i,st,\ldots,ed} - \left(\bar{E}_{i,.} \times F^T_{st,\ldots,ed,.}\right) \right\|$
    $\text{Err2} = \left\| B_{i,st,\ldots,ed} - \left[\left(\bar{L}_{i,.} \times F^T_{st,\ldots,st+t,.}\right), \left(\bar{H}_{i,.} \times F^T_{st+t+1,\ldots,ed,.}\right)\right] \right\|$
    if Err2 < Err1 **then**
      save $\bar{L}_{i,.}$ and $\bar{H}_{i,.}$ to $G^p_{i,.,.}$
      call recursive_splitting$(F, B, \text{size}, i, st, st + t, \bar{L}_{i,.})$
      call recursive_splitting$(F, B, \text{size}, i, st + 1 + 1, ed, \bar{H}_{i,.})$
      exit

Ensemble sliding window deconvolution (Step 1.3): For gene i, a sliding window is defined over the sorted list of samples with a specific window size s. For each window of sorted samples, problem (3) is solved using NNLS to estimate

the expected cell-type-specific expression across samples within that window. Cell-type-specific expression for each individual is then approximated by redistributing population-level estimates of cell-type-specific expression within each sliding window (This is again based on the assumption that subsets of individuals with very similar bulk expression profiles have a shared cell-type-specific expression profile. For more details on how this is done, please refer to the CIBERSORTx algorithm [163]). Thereafter, the initial approximate predictions from the sliding window deconvolution of window size s are refined using a linear-regression-based smoothing procedure such that the distribution of expression values is statistically consistent with population level estimates over each subset of patients from the p-freedom estimation step. This is based on the assumption that the estimated distribution of cell-type-specific expression in each subset is robust to outliers.

Given that this solution is a function of the window size, which is an artificially defined parameter, we suspect that a consensus solution obtained from averaging an ensemble of solutions from different window sizes would be more robust and closer to the ground truth. Hence, in our ensemble sliding window deconvolution, we set up window sizes ranging from $s_1 = \frac{1.5\times\text{number of cell types}}{0.8} \times \text{number of cell types}$ to $s_t = \max\left(4 \times \text{number of cell types}, \frac{\text{sample size}}{2}\right)$ and then perform the above sliding window deconvolution for each of these window sizes. Given multiple solutions for the cell-type-specific expression profile of each sample derived from multiple choices of sliding-window sizes (s_1 to s_t), their average is computed to obtain a single initial approximate solution to problem (1) (we refer to this as ensemble of window sizes). The above steps are repeated for the next gene until all the genes are done. Given

a re-implementation of the CIBERSORTx sliding window algorithm (as function sliding_window()), the ensemble sliding window pseudo code is provided below:

Algorithm 2 Ensemble Sliding Window

Input:
F= cell fraction matrix
B = bulk expression matrix
n = number of samples
i = gene index
G^p = expected cell-type-specific expression distribution over samples obtained from recursive splitting step
c = number of cell types
m = number of genes

Output:
Cell-type-specific (3-dimensional) expression matrix for all samples: G

procedure ENSEMBLE SLIDING WINDOW
    $s_1 \leftarrow 1.9 \times c$
    $s_t \leftarrow \max\left(4 \times c, \frac{n}{2}\right)$
    $G \leftarrow \text{Zeros}(m \times n \times c)$
    for w = s_1 to s_t **do**
        $G \leftarrow G + \text{sliding\_window}(B, F, G^p, w, i)$; (See CIBERSORTx algorithm for more details)
    $G \leftarrow \frac{G}{s_t - s_1 + 1}$
    return G

5.2.4.4 Confidence ranking of predictions from module 1

We expect that genes that follow the modeling assumptions of module 1 are more likely to have their cell-type-specific expression levels predicted confidently. Hence, while executing module 1, we collect a series of features that could be useful in determining confidence level of expression predictions for each gene-cell-type pair. These are: p-value of t-test determining if a gene is weakly expressed in a cell type (obtained from completion of step 1.1), ratio of mean predicted expression

levels and p-value of differential expression between subsets of samples h_1 and h_2 (obtained from completion of steps 1.2 and 1.3), Spearman correlation between predicted cell-type-specific expression and bulk gene expression across samples, Spearman correlation between bulk expression and the cell fraction across samples, etc. We then define a ranking Φ using this feature space such that genes achieving a high rank are on average ranked highly by each feature as follows:

$$\Phi(\text{gene i}, \text{cell type k}) = \frac{\sum_{\text{feature}\in\text{set}} \text{rank}(\text{feature}(\text{gene i}, \text{cell type k})}{|\text{set}|}$$

where Φ(gene i, cell type k) represents the prediction rank of gene i in cell type k, feature represents each feature we collected in the feature set, |set| represents the number of features and feature(gene i, cell type k) denotes each feature of gene i in cell type k. The values taken by each feature are arranged so that for features representing p-values, lower the value higher the rank, but for features representing Spearman correlations, higher the value higher the rank.

Additionally, it is well known that (the proteins encoded by) genes may interact with each other and behave collaboratively as complexes [100]; also, gene regulation is highly dependent on numerous regulatory elements including transcription factors [124, 120]. When looking at single cell expression data from three independent single cell datasets, we indeed find that expression profiles of 1000 randomly selected genes within the same cell type are much more strongly correlated than expected by random chance (Figure 5.4). Hence, we reason that genes with correlated expression

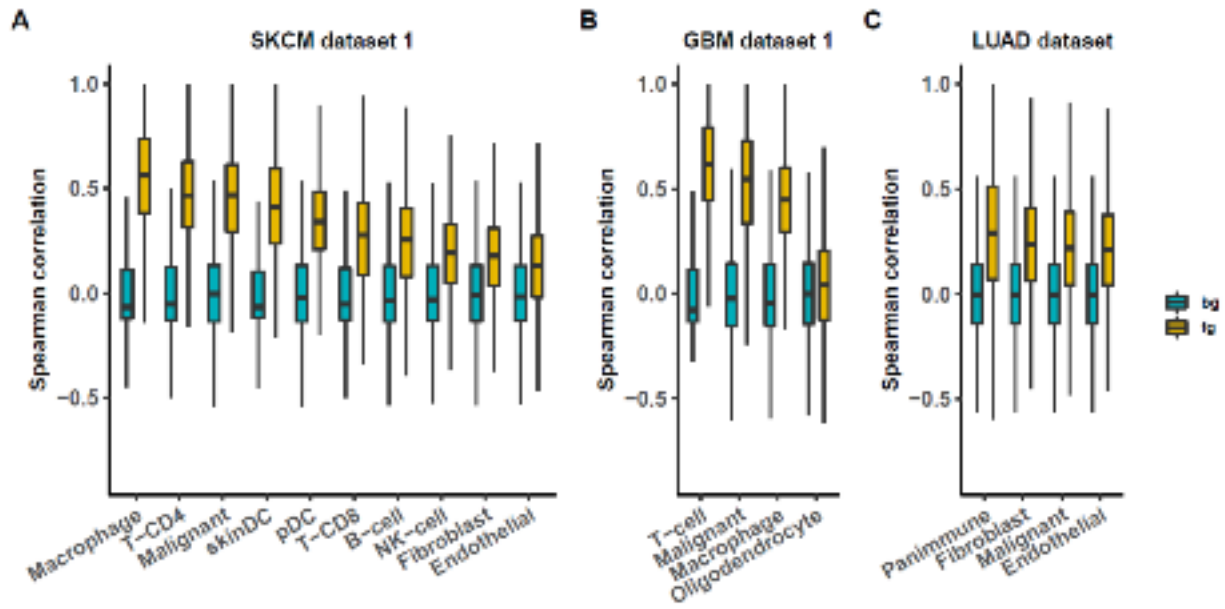

Figure 5.4: (A-C) boxplots depicting the gene-gene expression correlation distributions among cell types in SKCM dataset 1, GBM dataset 1 and LUAD dataset respectively for 1000 randomly selected genes. In each of the three plots, corresponding random-permutation-based background controls are provided. The yellow box represents the correlation derived from the original datasets as the foreground (fg), while the green box represents that derived from the randomly permuted background control (bg). The y-axis denotes the Spearman correlation value and the x-axis denotes the cell type.

predictions for a given a cell type will have similar confidence levels. Therefore, ranking Φ is updated to Φ^1 by accounting for these correlations as follows:

$$\Phi^1(\text{gene i}, \text{cell type k}) = \text{rank}_k(\max \Phi(\text{gene j}, \text{cell type k}) : j \in Q)$$

Here, Q represents the set of genes whose predicted expression in cell type k is strongly correlated with the predicted expression of gene i in cell type k (Spearman correlation ≥ 0.4).

For each cell type k, we define two disjoint but non-exhaustive subsets: $\mathcal{H}_k$ and $\mathcal{L}_k$, which we call the "high" and "low"-confidence sets of cell type k. Genes belonging to the set $\mathcal{L}_k$ will be passed on to module 2. Let m_k be the number of

genes whose predicted expression distribution in the population is at least bimodal (i.e., fold change in expression between the subsets h_1 and $h_2 > 1$). Genes are then assigned to the high, low confidence set of each cell type by the confidence ranking system using the following rule:

Add gene i to set:

- $\mathcal{H}_k$, if $\Phi^1(\text{gene i, cell type k}) < \text{round}\left(\frac{m_k^2}{2m}\right)$
- $\mathcal{L}_k$, if $\Phi^1(\text{gene i, cell type k}) > m - \text{round}\left(\frac{m_k 2}{2} \times \left(1 - \frac{m_k}{m}\right)\right)$

The results of the assignment are stored in a confidence matrix C (m genes $\times$ c cell types) encoding the high vs low confidence memberships of each gene in each cell type.

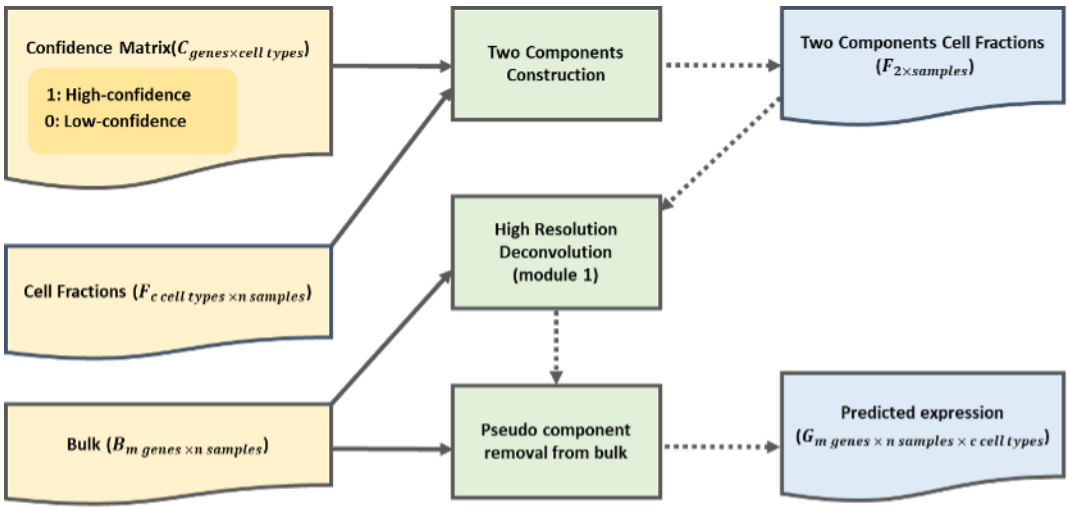

Figure 5.5: Given the estimated cell fractions F (c cell types $\times$ n samples), bulk and confidence levels estimated from module 1, for each cell type k we merge all the other cell types as a pseudo component to construct a two-component model. Thereafter for each low-confidence gene i in cell type k, we run module 1 to predict the expression in the pseudo component and finally remove the estimated expression of the pseudo component from the bulk to estimate the expression of the low-confidence gene i in cell type k.

5.2.4.5 Module 2 - Hierarchical deconvolution for low-confidence genes emerging from previous step

In this module, we simplify the general cell-type-mixture model described in module 1 to a 2-component mixture model (Figure 5.5). Specifically, for a gene i in the low-confidence set of cell type k, its observed bulk expression level in a sample is modeled as a mixture of 2 components: the first component represents the cell type k and the second component represents a pseudo-cell-type that is a composite of all the cell types except k^{th} cell type. We then re-run module 1 to predict individual specific expression of gene i for the pseudo cell-type. Finally, the prediction for the pseudo cell-type is subtracted from the bulk to approximately re-estimate the individual specific expression of gene i in cell type k. This is based on the assumption that the expression of the pseudo component might be better predicted than the expression of cell type k using module 1, especially if cell type k is not abundant or gene i is weakly expressed in cell type k. The above steps are repeated for all the remaining genes in the low-confidence set of each cell type. The hierarchical deconvolution pseudo code is provided below:

Algorithm 3 Hierarchical Deconvolution

Input:
F= cell fraction matrix
B = bulk expression matrix
C = confidence level matrix (records low confidence genes in each cell type that need to be re-evaluated by module 2)
G = predicted gene expression in each cell type and sample (Output of module 1)
c = number of cell types
m = number of genes

Output:
Updated predictions of gene expression in each cell type and sample in G

procedure HIERARCHICAL DECONVOLUTION
  for k = 1 to c **do**
    $F'' \leftarrow [F[k,], 1 - F[k,]]]$
    for i = 1 to m **do**
      if $C[i,k] = 0$ **then**
        $G^{\text{pseudo}} \leftarrow$ High_resolution _deconvolution(B, F'', i)
        $G[i,,k] \leftarrow \frac{B[i,] - F''[2,] \times G^{\text{pseudo}}[i,,2]}{F''[1,]}$

5.2.4.6 Confidence ranking of predictions emerging from module 2

Following module 2, we re-rank all genes in the low confidence set L_k of each cell type by re-defining the ranking Φ^2 as follows:

For gene $i \in \mathcal{L}_k$,

$$\Phi^2(\text{gene i}, \text{cell type k}) = \text{rank}_k \left(\frac{1}{|\mathcal{H}_k|} \sum_{j \in \mathcal{H}_k} \rho(\text{gene i,gene j}) \right)$$

Where ρ(gene i,gene j) represents the Spearman correlation between new predictions of gene i and old predictions of gene j, and $|\mathcal{H}_k|$ represents the number of genes in high confidence set $\mathcal{H}_k$. This is again based on observations of single cell expression data described above from which we deduce that genes with similar

confidence levels are expected to have correlated predictions (Figure 5.4). We now describe how the confidence ranking system takes this new ranking of genes in the low confidence set of each cell type and decides which genes need to be upgraded to the high confidence set.

Let $|\mathcal{L}_k|$ be the number of genes in the low confidence set of cell type k, m be the total number of genes and CFM_k be the mean cell fraction of cell type k. The confidence ranking system upgrades the membership of genes from the low confidence set $\mathcal{L}_k$ to the high confidence set $\mathcal{H}_k$ using the following rule:

For gene $i \in \mathcal{L}_k$

- $\mathcal{H}_k \leftarrow \mathcal{H}_k \cup \{i\}$, if $\Phi^2(\text{gene i}, \text{cell type k}) < \text{round}\left(\frac{|\mathcal{L}_k|^2 \times CFM_k}{2m}\right)$
- $\mathcal{L}_k \leftarrow \mathcal{L}_k \cup \{i\}$, if $\Phi^2(\text{gene i}, \text{cell type k}) > m - \text{round}\left(\frac{|\mathcal{L}_k|^2}{2m}\right)$

The results of the assignment are stored in the confidence matrix C (m genes $\times$ c cell types) encoding the high vs low confidence memberships of each gene in each cell type.

5.2.4.7 Module 3 – Imputation-based deconvolution for low-confidence genes emerging from previous step

Module 3 operates on the assumption that the expression levels of two genes are supposed to be correlated in some cell types if we observe that their bulk expression is significantly correlated 16. For a gene i still in the low-confidence set of cell type k, the Spearman correlations between the bulk expression profile of gene i and bulk expression profiles of genes in the high confidence set of cell-type k are estimated.

If the bulk expression profile of gene i is highly correlated (Spearman correlation 0.5) with the bulk expression profiles of more than two genes in the high-confidence set of cell type k, then a lasso regression-based machine learning model is trained using bulk expression to impute individual specific expression of gene i in cell type k based on predicted expression profiles of high-confidence genes in cell type k (Figure 5.6). The above steps are repeated for all the remaining genes in the low-confidence set of each cell type.

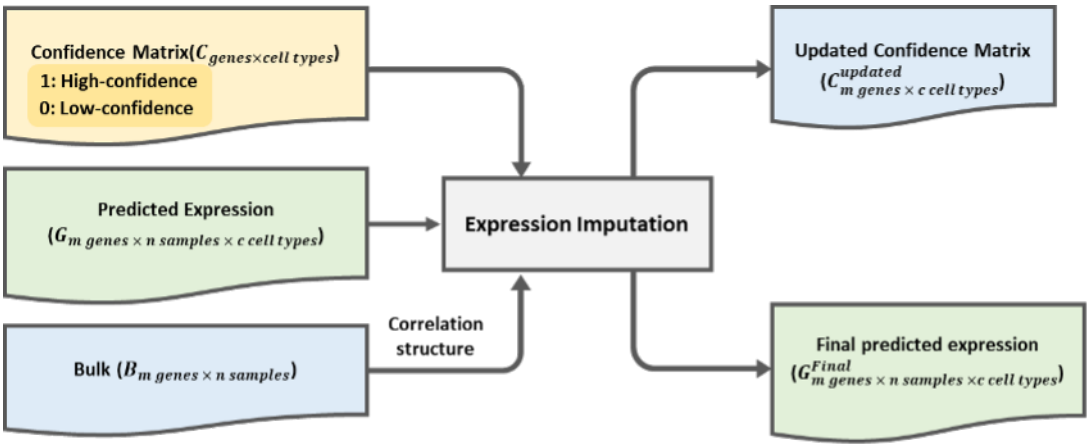

Figure 5.6: Given the predicted cell-type-specific expression G (m genes $\times$ n samples $\times$ c cell types), bulk and confidence levels estimated from module 1, in each cell type k and for each low-confidence gene i, we compute the correlation between gene i and each of other genes in bulk. If the number of genes which are highly correlated with gene i is more than 2, we build up a machine learning model to predict the expression of gene i in cell type k based on the expression of other high-confidence genes which are correlated with gene i. After imputation, both the predicted expression matrix G and confidence matrix C will be updated to record the final low/high confidence memberships of genes in each cell type.

Algorithm 4 Imputation Based Deconvolution

Input:
F= cell fraction matrix
B = bulk expression matrix
C = confidence level matrix (records low confidence genes in each cell type that need to be re-evaluated by module 3)
G = predicted gene expression in each cell type and sample (Output of module 2)
c = number of cell types

Output:
Updated predictions of gene expression in each cell type and sample in G

procedure IMPUTATION BASED DECONVOLUTION
  for k = 1 to c **do**
    $\mathcal{H}_k \leftarrow$ gene $j : C[j,k] = 1$
    $\mathcal{L}_k \leftarrow$ gene $j : C[j,k] = 0$
    $\text{Conf}_{\text{high}} \leftarrow$ all genes $\in \mathcal{H}_k$
    $\text{Conf}_{\text{low}} \leftarrow$ all genes $\in \mathcal{L}_k$
    $\text{Corrs} \leftarrow$ Spearman correlation matrix$(B[\text{Conf}_{\text{low}},], B[\text{Conf}_{\text{high}},])$
    $F'' = [F[k,], 1 - F[k,]]]$
    for each gene $i \in \mathcal{L}_k$ **do**
      if $\left(\sum_{j \in \mathcal{H}_k} \text{Corrs}[i,j] \geq 0.5\right) \geq 2$ **then**
        train imputation model: $B[i,]\ f_{\text{imp}}\left(B[\text{Conf}_{\text{high}},]\right)$
        impute $G[i,,k] \leftarrow f_{\text{imp}}\left(G[\text{Conf}_{\text{high}},,k]\right)$

5.2.4.8 Confidence ranking for predictions emerging from module 3

Following module 3, we collect the following confidence ranking features for each gene i in the low confidence set of cell type k: the correlations of predicted gene expression with bulk expression, the correlation between cell fractions and bulk expression, number of genes as features in the imputation model, average Spearman correlation between new predictions of gene i and predictions of genes in the high confidence set of cell type k. We re-define a ranking Φ^3 over all genes in the low confidence set of each cell type using this feature space such that genes achieving

a high rank are on average ranked highly by each feature. Genes that are ranked among top 80% (an artificial cutoff) of all genes in the low confidence set of a cell type k are now upgraded to the high confidence set of cell type k by the confidence ranking system.

5.2.4.9 Final output – confidence scores and cell-type-specific gene expression profiles of each sample

To transform high- vs low-confidence set memberships of genes in each cell type (which were based on artificially defined rules/cut-offs for easy implementation of the greedy algorithm), into scores that are continuous in the range $[0, 1]$, the following final steps were taken: (a) The pair-wise correlations between the predicted expression profile of a gene i in cell type k and predicted expression profiles of genes belonging to the high-confidence set of cell type k are averaged to generate a score for gene i; (b) the cell-type-specific expression predictions across the samples (columns) are randomly shuffled to generate a background and step (a) is repeated to estimate a background distribution of scores for each gene; (c) for each gene and each cell type, one can now determine an empirical p-value pv based on this background distribution of scores. These p-values quantify the probability of a gene having high confidence predictions by random chance if its predictions are correlated with predictions of any other genes belonging to the high-confidence set of a cell type. The p-values are low for genes that are part of the high confidence set and high for genes part of low confidence set and intermediate for genes belonging to neither. Hence,

we record $1 - pv$ as the final confidence score for each gene-cell-type pair. The final outputs of CODEFACS are the approximate solution for 3-dimensional matrix G after execution of module 3 (imputation-based deconvolution) and confidence scores for each gene-cell-type pair.

5.2.5 Inference of clinically relevant cellular crosstalk in the TME

In this section we describe the database of putative ligand receptor interactions between various immune cell types: LIRICS (LIgand Receptor Interactions between Cell Subsets) and its application to discover clinically relevant cellular immune crosstalk.

5.2.5.1 Curation of established ligand-receptor protein-protein interactions between cell types in the tissue microenvironment

Known protein-protein interactions between tumor, epithelial, immune and stromal cell types in the tissue microenvironment were manually curated from various resources. Specifically, interactions corresponding to cytokine/chemokine - cytokine/chemokine receptor interactions, ligand-receptor interactions involved in cell adhesion/leukocyte trans-endothelial migration, ligand-receptor interactions involving the TNF receptor superfamily and lastly, ligand receptor interactions involved in regulation of NK and T cell cytotoxicity were all merged into one Excel spreadsheet [35, 41, 110, 111, 159, 174, 242]. In total, 369 putative ligand-receptor interactions were collected. This list primarily covers proteins that have well characterized im-

munological functions. Certain receptors are complexes encoded by more than one gene, such as TGF beta family receptors. They are documented as a list of genes separated by a ";". Furthermore, certain proteins serve as both ligands on some cell types and receptors on other cell types, such as HVEM (TNFRSF14).

5.2.5.2 Expected distribution of ligands and receptors across different cell types from prior knowledge

The database assigns a binary indicator (1/0), for each ligand/receptor, across the compendium of cell types indicating with 1 if the ligand/receptor can be produced by a cell type based on prior evidence of cell-surface protein expression or secretion (0 otherwise). This knowledge was extracted from Appendix II-IV of Janeway's Immunobiology 9th Edition Textbook [159]. The appendix, in addition, records ligands/receptors whose expected cell type specific distribution is less precisely defined. For instance, certain cytokines/chemokines are reported to be broadly produced by lymphocytes. Hence, without additional evidence, it is reasonable to expect that such ligands/receptors can also be produced in specific contexts by all cell types that are lymphocytes. We formalize this notion by defining a functional equivalence class for each cell type. For instance, the functional equivalence class for B cells is defined as: lymphocytes, lymphoid cells, leukocytes, antigen presenting cells, nucleated cells, all cells. A schema representing such relationships is stored in the database. We then describe in a subsequent section how this prior knowledge can be used to systematically enumerate all ligands/receptors that can potentially

be produced by a specific cell type of interest.

5.2.5.3 Annotation of functional effects of ligand-receptor interactions on participating cell types

Certain ligand-receptor interactions between immune cell types have an activating or inhibitory effect on the cell type expressing the receptor (also known as the target cell type) or in some cases both the ligand and receptor expressing cell types (regarded in literature as costimulatory). Discovery of such interactions resulted in the development of immune checkpoint blockade therapy such as anti-PD1 and anti-CTLA4 which has revolutionized cancer treatment. We systematically curated literature on all such interactions from [35, 41, 174, 242, 23, 219, 80, 235, 220, 40] and classified them into two ontologies as follows:

- **Activating/costimulatory** encapsulates interactions with the following functional characteristics reported in literature: increased cytotoxicity, increased cytokine production, increased cell proliferation, increased cell survival, existence of immunoreceptor tyrosine-based activation motifs (ITAMs) in the cytoplasmic tail of the receptor.

- **Inhibitory/checkpoint** encapsulates the following functional characteristics reported in literature: decreased cytotoxicity, exhaustion, reduced cytokine production, decreased TCR signaling activity (for T cells), reduced cell proliferation, reduced cell survival, existence of Immunoreceptor tyrosine-based inhibitory motifs (ITIMs) in the cytoplasmic tail of the receptor.

In the database, interactions with conflicting effects reported on target cell types or for cases where the effect of the interaction depends on other factors were left un-annotated. In addition to activating/inhibitory interactions, the database also annotates other interactions based on prior knowledge from Janeway's immunobiology 9th Edition Textbook [159].

- **Pro-inflammatory** interactions involving inflammation mediator cytokines such Interferon Gamma, TNF-alpha, IL1, IL12 and IL18
- **Chemotaxis** cytokine/chemokine interactions involved in cell chemotaxis in regular or inflammatory conditions (responsible for lymphocyte infiltration)
- **Cell-adhesion** interactions involved in cell adhesion/leukocyte trans-endothelial migration (responsible for extravasation from blood vessels to tissue)

5.2.5.4 LIRICS STEP 1: Querying all plausible ligand receptor interactions between any two cell types based on prior knowledge

In this step, we query all ligand receptor interactions that could potentially take place between two cell types A and B. The user can plug in the names of any two cell types whose names match with the names of cell types in the database and then LIRICS lists all ligand-receptor interactions that could potentially take place between cell types A and B. This list is determined by first finding which ligands/receptors can potentially be produced by each cell type (cell type A and B) based on prior knowledge of the expected distribution of ligands and receptors on

cell types. It then adds to this list any ligands/receptors that are expected to be found in the functional equivalent cell types. Given a set of all potential ligands and receptors on each cell type, LIRICS returns all known physical protein-protein interactions involving these ligands and receptors.

5.2.5.5 LIRICS STEP 2: Identifying which plausible interactions are likely to occur (or "active") in each sample given deconvolved gene expression data from CODEFACS

Given a queried set of all plausible ligand receptor interactions between cell types $(A, B, C, ...)$: $\{(L_1^A, R_1^B), (L_2^A, R_2^C), \ldots, (L_1^C, R_1^B), \ldots, \}$, one can integrate this prior knowledge with deconvolved expression data from CODEFACS to infer which interactions are likely to occur in each sample as follows:

For any two cell types A and B with a plausible ligand-receptor interaction (L_z^A, R_z^B), we define a binary indicator $Z_{L_z^A, R_z^B} \in \{0, 1\}$, such that $I_{L_z^A, R_z^B} = 1$ if a physical interaction between (L_z^A, R_z^B) is likely to take place in a sample, and has the value 0 otherwise. An interaction is considered likely to take place (synonym: "active") in a sample if the ligand L_z^A is overexpressed in cell type A and receptor R_z^B is over expressed in cell type B, in that sample. To determine if ligand L_z^A and receptor R_z^B are over-expressed in cell types A and B in a given sample, we use the median deconvolved expression of the ligand L_z^A in cell type A over all input samples and likewise the median deconvolved expression of receptor R_z^B in cell type B over all input samples as controls. Ligands such as cytokines and chemokines, can be

secreted by cells and hence are not surface bound. However, we expect the levels of secreted cytokines/chemokines by a cell type to be proportional to their cell-type-specific gene expression. Furthermore, multiple genes are required to encode certain ligands/receptors, each gene being part of a specific subunit in the protein complex. For such ligands/receptors to be expressed, all genes required to build the ligand or receptor need to be expressed. Hence, we assume the expression of such receptors or ligands in a cell type is the minimum of the expression of individual genes constituting the ligand or receptor.

This approach has two key advantages, besides being biologically intuitive. First, the binary indicator is expected to be robust to noise in gene expression despite the varying levels of confidence in the predicted cell-type-specific gene expression from different datasets. This follows from the statistical properties of median-based filters in signal processing [251]. Second, it enables comparison of individual profiles independent of the dataset source due to their shared biological representation if the datasets being compared have expression measurements of the same genes. Thus, one can seamlessly pool multiple datasets together, augment sample size and increase statistical power.

5.2.5.6 LIRICS STEP 3: Downstream enrichment analysis and visualization

Given this binarized representation, one can perform a Fisher's exact test to assess if any specific cell-cell interaction is more likely to occur in samples with a

specific phenotype compared to a control group. This is quantified by computing the enrichment score, expressed as an odds ratio of the interaction in each phenotype of interest. A score around 1 indicates a neutral trend, a score >1 indicates enrichment of the interaction in the phenotype of interest and a score close to 0 indicates enrichment in the control group. Furthermore, the associated p-values of each test can be inspected post multiple hypothesis testing correction to identify any significant trends in the data. One can also plot the most significant trends occurring in a network where each edge represents a ligand-receptor interaction between two cell types and the thickness of the edge is proportional to the enrichment score of the interaction in a phenotype of interest. The circlize package in R is used to make these plots [84].

5.2.6 Feature selection and machine learning

We used a genetic algorithm, which is a randomized heuristic search algorithm designed to select optimal features for a prediction task given some user-defined fitness function for training [146]. In this setting, the features are ligand-receptor interactions between cell-types, the prediction task is predicting response to ICB treatment and the fitness function is defined as the accuracy of predicting a user defined phenotype based on the total number of interactions from those selected occurring in a given sample; accuracy is quantified by the AUC. To reduce the risk of over-fitting and aid in faster convergence of the genetic algorithm during training, the size of the search space is reduced by first removing any ligand-receptor

interactions with multi-gene receptors or ligands. These features are expected to be noisy because the relationship between expression of genes encoding the protein complex and the cell surface expression of the protein complexes is less well defined. Second we assess the fold change of each feature between the two classes specified in the training dataset (e.g., hypermutated vs non-hypermutated) and only select those with a fold change >1. These features are then passed to the genetic algorithm for further optimization.

The algorithm starts out by randomly generating sets of features. This is defined as the seed population. These sets iteratively evolve via the phenomenon of natural selection enforced by the user defined fitness function. Specifically, for each subsequent iteration, features from the best performing sets, as determined by the user defined fitness function, in the current iteration are mixed at random followed by random new feature additions or dropouts (referred to as mutations) to build a new generation of feature sets and the process repeats. Eventually, after a number of epochs, which we set to 100, the fitness function converges to an optimum and the best set of features for the prediction task is returned to the user. Since the fitness function landscape is often non-convex and the training process is stochastic, we repeat the training process 500 times, each with a randomly chosen seed population, and eventually choose frequently selected features over all solutions to reach a solution we suspect is close to the global optimum solution. For our plots, we set the threshold to frequency > 100 times. The probability of any feature being selected more than 100 times by random chance based on this approach is estimated to be < 0.01. Results are qualitatively similar for more stringent thresholds. The genetic

algorithm was implemented in R using the genalg package [246].

5.3 Results

5.3.1 Overview of CODEFACS and LIRICS

CODEFACS is designed to characterize the tumor microenvironment by reconstructing the cell-type-specific transcriptomes of each sample from bulk expression. It takes as input the bulk RNA-seq expression values of a cohort of tumor samples and either the estimations of the cell fractions of a pre-defined set of cell types in each sample or their cell-type-specific molecular signature profiles, derived based on reference datasets or from the literature [163].

CODEFACS then employs a heuristic approach that sequentially executes three modules: (module 1) high resolution deconvolution, (module 2) hierarchical deconvolution and (module 3) imputation. Each module is designed to predict the cell-type-specific expression of genes in each sample; the second and third modules aim to overcome the shortcomings of the previous modules. A key component of CODEFACS is its confidence ranking system, which receives cell-type-specific expression predictions from the different modules and labels them as high or low-confidence estimations. Genes whose expression is determined with high confidence in a given module are added to the output set, while low confidence predictions are continued to be processed in subsequent modules (See Figure 5.7 panel A). The final output of CODEFACS consists of two items: (a) a three-dimensional gene expression matrix, where each entry represents the predicted gene expression a gene

in a given cell-type in a specific sample, and (b) a two-dimensional matrix of confidence scores ranging from [0,1] representing which gene-cell-type pairs have most confident predictions (1) and which pairs have least confident predictions (≈0; See Figure 5.7 A, Output). These scores can be further investigated to assess the quality of predictions for a given dataset.

Given fully deconvolved gene expression data from CODEFACS, one can use LIRICS (LIgand Receptor Interactions between Cell Subsets) (Figure 5.7, panel B) to transform this data into a biologically interpretable feature space of active ligand-receptor interactions between cell types in each sample. Specifically, LIRICS takes the output of CODEFACS and processes it in three steps: (step 1) the first step queries a database of all plausible ligand-receptor interactions between any two cell types A and B, that we have systematically assembled and curated from the literature. This database is publicly available as part of LIRICS. (step 2) In the second step, given the deconvolved expression profiles of cell type A and cell type B in a given bulk tumor sample, LIRICS denotes as 'active' or 'likely to occur' ('1') the interactions where both the ligand and receptor are over-expressed in the relevant cell-types in that sample, or otherwise 'inactive' ('0'). A ligand or receptor is considered to be over-expressed in a given cell type if its expression exceeds the median expression in that cell type (Supplementary Note). (step 3) Finally, a Fisher's enrichment analysis is performed to test the association of the activity of specific ligand-receptor interactions with any relevant phenotypes of interest (e.g., treatment response, mutational subtype, etc.) (Figure 5.7, panel B). Furthermore, if required, one can collectively analyze the binary profiles returned by LIRICS

from multiple independent datasets to augment sample size and increase statistical power. Finally, one can apply a false discovery rate (FDR) cut-off post multiple hypothesis correction and visualize significantly enriched interactions in a volcano plot as shown at the bottom of Figure 5.7, panel B, or, alternatively, display their network structure as shown in subsequent biological applications.

5.3.2 Benchmarking CODEFACS performance

To assess the accuracy of CODEFACS, we generated 15 benchmark datasets (see Methods) by merging publicly available single cell RNA-seq [103, 74] and FACS sorted purified RNA-seq [72]. Thereafter, we applied CODEFACS to deconvolve these generated bulk datasets and define the accuracy of its predictions by computing the Kendall correlation between the predicted and ground truth expression in each cell type across individual samples (the Kendall correlation provides a less inflated measure of accuracy by accounting for ties in the data). In the main text, we show the results obtained on three benchmark bulk datasets: one derived form a FACS-sorted lung cancer data, one from a single cell melanoma RNA-seq data and from a single cell glioblastoma RNA-seq dataset, respectively. Each bulk sample from these datasets represents a real biopsy from a patient. We show that CODEFACS can predict the cell-type-specific expression of more genes than CIBERSORTx (with Kendal's correlation ≥ 0.3) (Figure 5.8 panels A,B,C), and its predictions are overall more accurate (Figure 5.8, panels D,E,F). The results for all the remaining 12 benchmark datasets, created via artificial mixing of single cell profiles and single

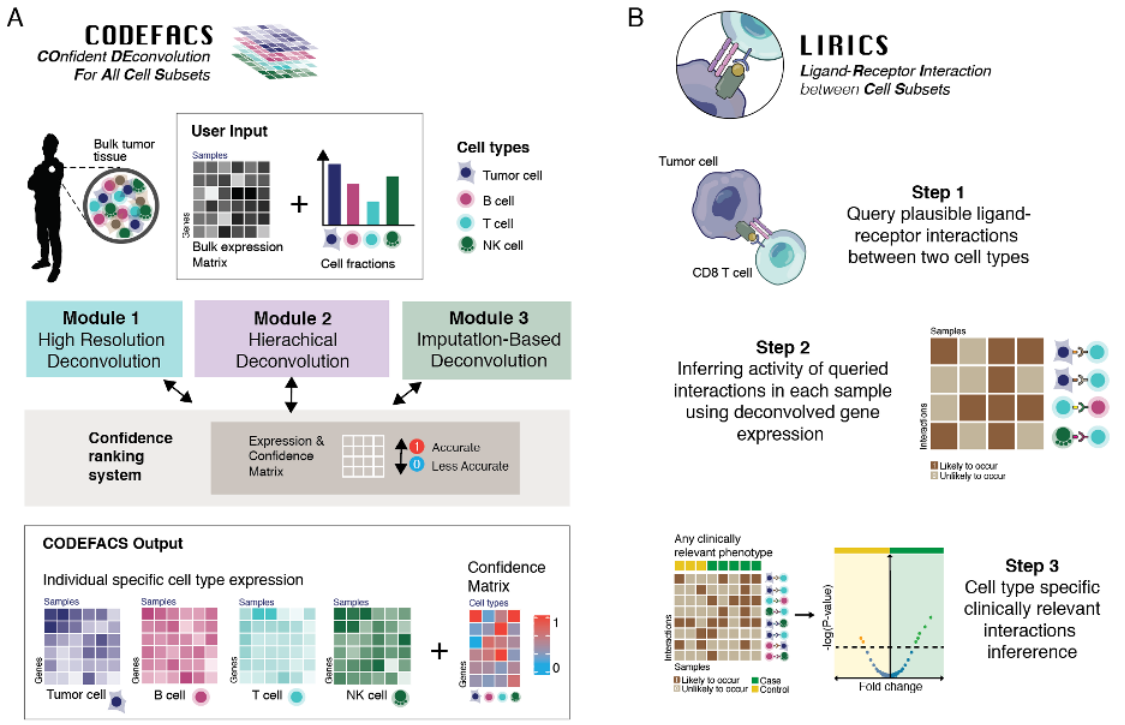

Figure 5.7: CODEFACS takes bulk gene expression profiles and prior knowledge of the cellular composition of each sample and executes a heuristic three step procedure to infer the deconvolved gene expression in each sample. In module 1, we perform a high-resolution deconvolution, which extends the CIBERSORTx algorithm. In module 2 (hierarchical deconvolution), bulk expression is modeled as a mixture of two components: a specific cell type of interest and all the remaining cell types. The expression for the cell type of interest is predicted by removing the estimated expression in the second component (using high-resolution deconvolution from module 1) from the bulk mixture. In module 3 – imputation-based deconvolution, we impute the cell-type-specific expression of a specific gene based on the predicted cell-type-specific expression of other high-confidence genes that are co-expressed with that gene in the bulk. Each module is designed to overcome the shortcomings of its predecessor based on their respective modeling assumptions. The confidence ranking system is responsible for classifying all the predictions at the end of each module into high-confidence or low-confidence predictions. Genes classified into the low-confidence class at the end of one module (e.g. module 1) are passed to the next module (e.g. module 2) for refinement. Finally, after all the three modules are executed, the prediction confidence levels are re-evaluated. The final output of CODEFACS consists of a 3-dimensional matrix with cell-type-specific gene expression predictions for each sample, along with estimated confidence scores of predictions for each gene in each cell type. For more details, see Supplementary note. (B) LIRICS takes the output of CODEFACS and processes it in three steps. In step 1, for each possible permutation of cell type pairs, LIRICS queries a literature-curated repository for enumerating all plausible ligand-receptor interactions between specified cell types. In step 2 this prior knowledge is integrated with the output from CODEFACS to infer which of the plausible cell-cell interactions are likely to occur or be "active" in each individual sample. The result is a binary matrix with rows representing each plausible cell-cell interaction and columns representing each patient's tumor sample. Finally, in step 3, given any clinically relevant phenotype (e.g. response to therapy, driver mutation status, etc..), one can perform a Fisher's enrichment analysis (shown at the bottom) to discover cell-grounded receptor-ligand interactions in the TME that are associated with the phenotype of interest.

cell RNA-Seq imputation, also show the superiority of CODEFACS (supplementary Figure 5.9 and 5.10). Overall, we observe that the more abundant the cell type is, the better CODEFACS can predict its cell-type-specific gene expression (Figure 5.11).

Next, we quantified how well the confidence scores it returns align with the Kendall scores that measure the true prediction accuracy with the ground truth for each (gene, cell-type) pair to validate the claim that our confidence scores can help the user filter out potentially noisy predictions in real bulk datasets. We quantified this using two metrics: Spearman correlation and a classification AUC (Area Under the ROC Curve). Across all the benchmark datasets analyzed, the Spearman correlation between confidence scores of genes in each cell-type and their corresponding Kendall scores (quantifying the true prediction accuracy) is strong and positive (Figure 5.8, panel G depicts the results for the FACS sorted lung cancer benchmark dataset, and Figure 5.12 depicts the results for the remaining benchmark datasets); To perform a classification-based quantification, we grouped the genes in each cell-type into two classes based on the correlation between their predicted and actual expression, informative (prediction accuracy ≥ 0.1 and p-value ≤ 0.05) and uninformative (prediction accuracy < 0.1 or p-value > 0.05). We then tested whether the confidence scores could be used to classify genes into these two classes for each cell type. We find that the confidence score could effectively filter out uninformative predictions (Figure 5.8, panel H depicts the results for the FACS sorted lung cancer benchmark dataset, and supplementary Figure 5.13, depicts for the remaining benchmark datasets).

Finally, to further evaluate CODEFACS on real bulk tumor data (where the ground truth is unavailable), we applied it to deconvolve bulk expression data from 21 cancer types (8000 RNA-seq samples) in TCGA. To infer the cellular abundance of each cell type in each sample which is required as input for CODEFACS, we made use of matched bulk methylation data available for these samples and methylation-based reference signature profiles of distinct cell types. These include 11 cell type signatures (macrophages/dendritic cell:CD14+, B cells: CD19+, CD4+T cells, CD8+ T cells, T reg cells, NK cells: CD56+, endothelial cells, fibroblasts, neutrophils, basophils, eosinophils and tissue-specific tumor cells) obtained from MethylCIBERSORT [39]. Reassuringly, we found strong Spearman correlations between the resulting predicted tumor cell fraction and the tumor purity estimates derived from matched mutation and copy number data (based on ABSOLUTE) for the same samples across 10 cancer types (Spearman correlation: min=0.72, max=0.88, avg=0.8). This testifies that methylation-based cell fraction estimates indeed form a reliable basis for running CODEFACS to deconvolve TCGA samples.

We then asked if CODEFACS can recover the expected cell-type-specific gene expression signature of different cell types in a given cancer type. To this end, we computed the Spearman correlation between (a) the mean deconvolved gene expression of the top confidently deconvolved genes in a given cell type (confidence score $\geq$ 0.95) and (b) the mean expression of these genes, which we derived from completely independent single cell expression data of the same cancer type (Methods). We find that (a) and (b) are substantially correlated (Figure 5.8, panel I depicts results for the TCGA-LUAD (lung adenocarcinoma) dataset as an example and supplementary

Figure 5.14 for the remaining cancer types that have publicly available scRNA-seq data). The concordance level is higher for cell types that are abundant (e.g., tumor cells and fibroblasts) and decreases for less abundant cell types. Additionally, we observed that tumor cells have the largest fraction of genes whose expression is predicted with high confidence, with the highest in thyroid cancer (THCA, 67.4% of all genes). Furthermore, 7 KEGG pathways are significantly enriched (adjusted p-value < 0.01) with highly confident genes in tumor cells (confidence score ≥ 0.95) across the 21 cancer types. Those pathways mostly involve RNA transport, spliceosome, DNA replication, and mismatch repair.

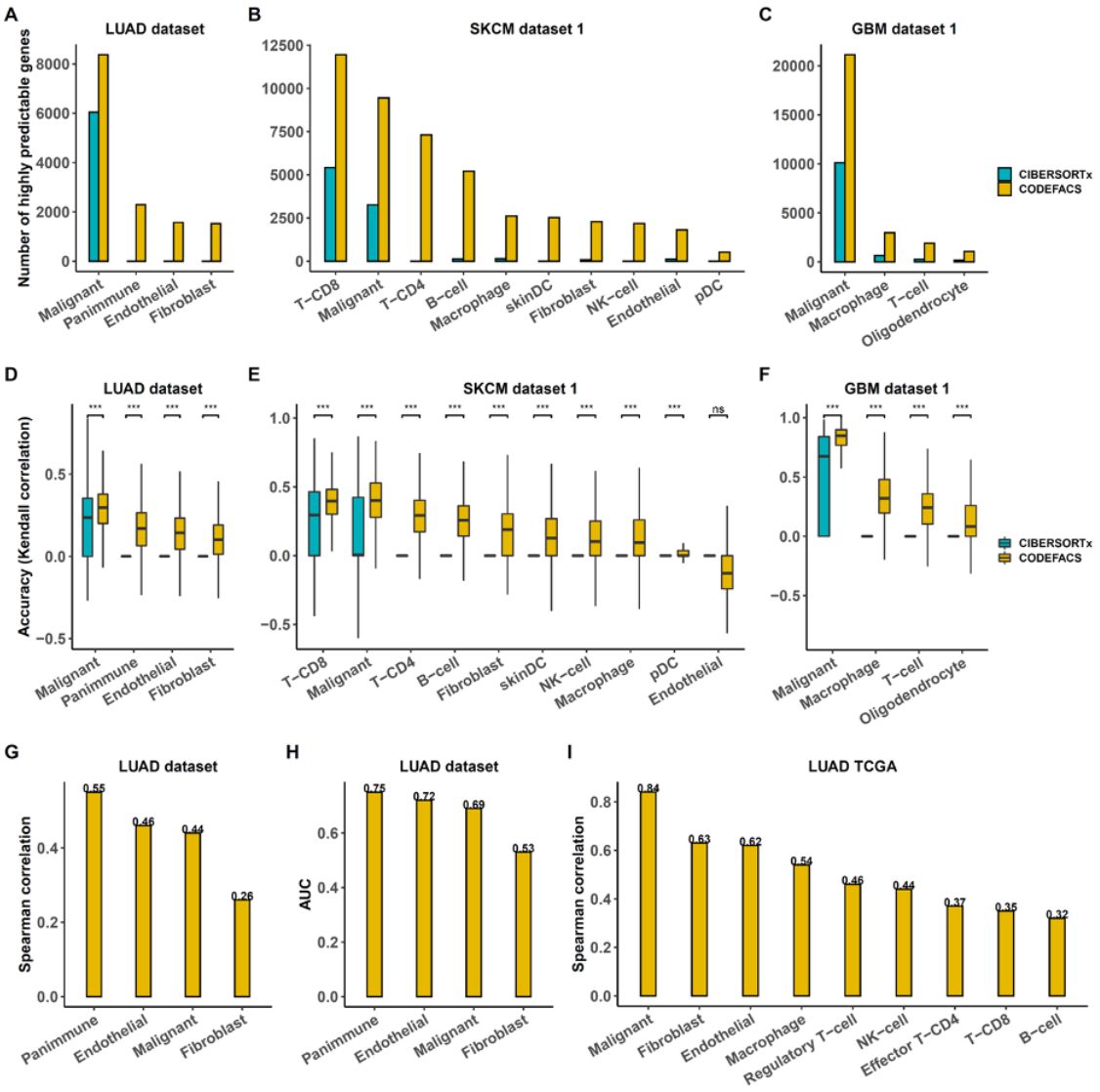

Figure 5.8: (A-C) bar plots depicting the number of genes with a prediction accuracy (Kendall correlation) ≥ 0.3 with the ground truth for each cell type, estimated from bulk-generated samples of lung cancer (LUAD dataset; sample size = 26) [72], melanoma (SKCM dataset 1; sample size = 28) [103] and glioblastoma (GBM dataset 1; sample size = 24) [74] benchmark datasets, as estimated by CODEFACS (yellow bars) and CIBERSORTx (blue bars). (D-F) boxplots depicting prediction accuracy distributions of all genes across different cell types in the lung cancer (LUAD with sample size 26) [72], melanoma (SKCM with sample size 28) [103] and glioblastoma (GBM with sample size 24) [74] benchmark datasets, using CODEFACS (yellow) and CIBERSORTx (blue). A two-sided Wilcoxon signed rank test was performed to compare the prediction accuracies of CODEFACS and that of CIBERSORTx for each cell type in each dataset. $\star\star\star$ denotes p-values $<$ 2e-16. (G) Spearman correlations between prediction accuracies and confidence scores among cell types in the lung cancer benchmark dataset (LUAD dataset; sample size = 26) [72]. The y-axis indicates the spearman correlation coefficient value, while the x-axis indicates the cell type. (H) AUCs obtained in classifying informative and uninformative predictions among cell types in lung cancer benchmark dataset (LUAD dataset; sample size = 26) [72]. (I) bar plots depicting the Spearman correlations between mean deconvolved cell-type-specific expression in TCGA-LUAD and mean cell-type-specific expression derived from publicly available single cell datasets of LUAD. The y-axis indicates the Spearman correlation coefficient value, while the x-axis indicates the cell type.

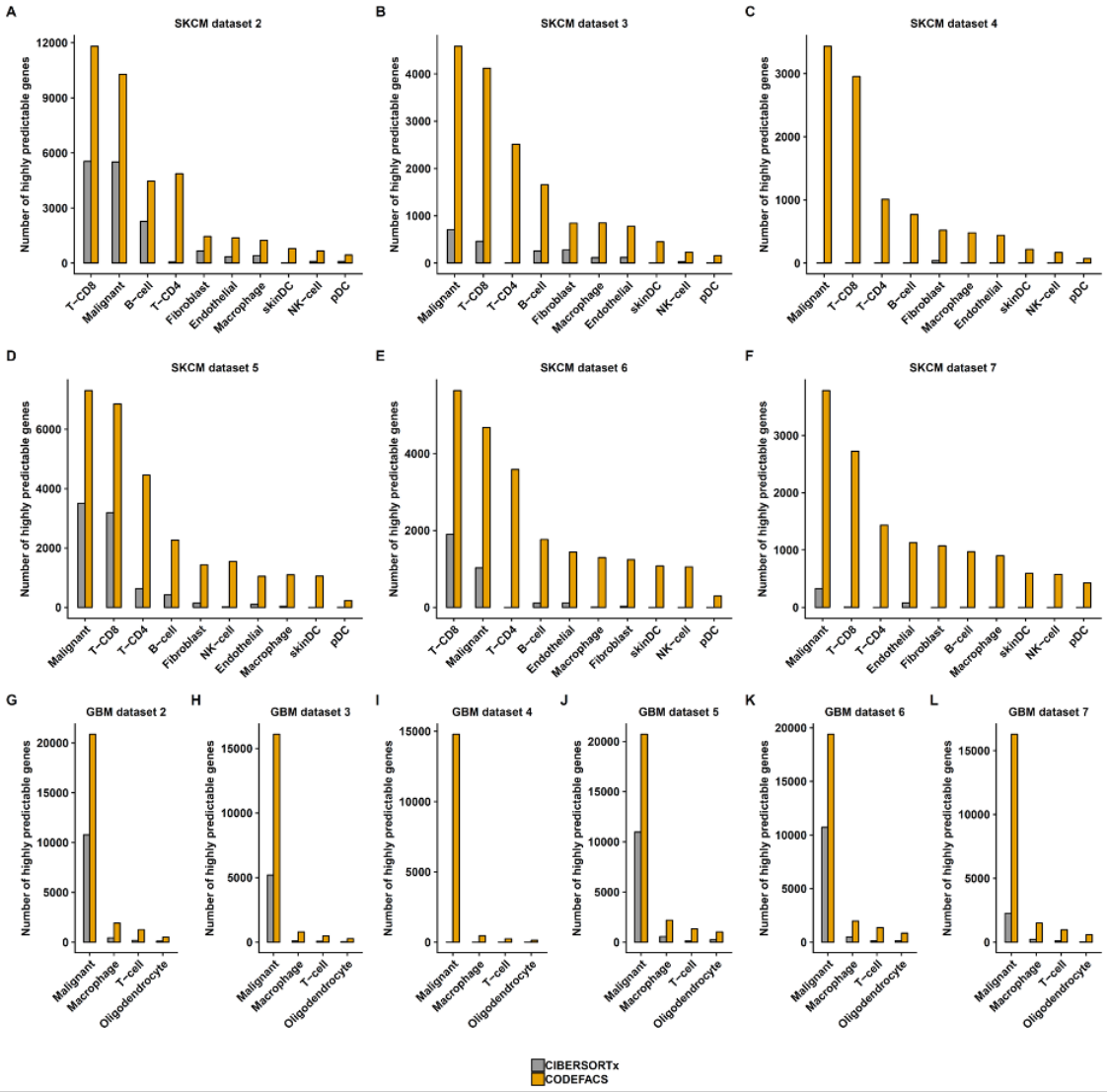

Figure 5.9: (A-L) bar plots depicting the number of highly predictable genes (Kendall correlation ≥ 0.3) among 12 validation datasets (SKCM dataset 2, SKCM dataset 3, SKCM dataset 4, SKCM dataset 5, SKCM dataset 6, SKCM dataset 7, GBM dataset 2, GBM dataset 3, GBM dataset 4, GBM dataset 5, GBM dataset 6, GBM dataset 7).The yellow bar represents the performance of CODEFACS, while the gray bar represents that of CIBERSORTx.

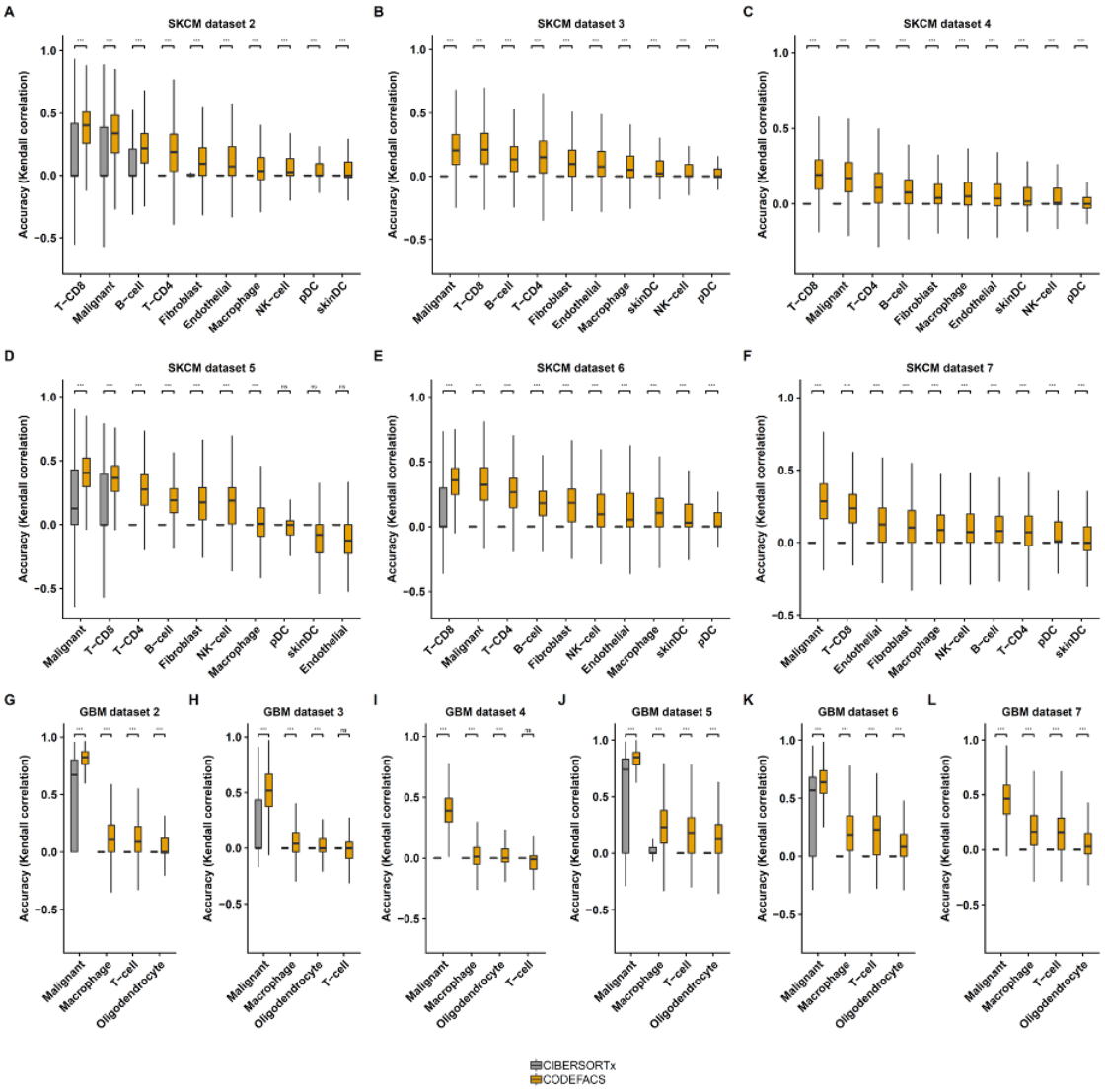

Figure 5.10: (A-L) boxplots depicting the accuracy distributions among the 12 benchmark dataset (SKCM dataset 2, SKCM dataset 3, SKCM dataset 4, SKCM dataset 5, SKCM dataset 6, SKCM dataset 7, GBM dataset 2, GBM dataset 3, GBM dataset 4, GBM dataset 5, GBM dataset 6, GBM dataset 7). The yellow boxes represents the performance of CODEFACS, while the gray boxes represents that of CIBERSORTx. Wilcoxon signed rank test was performed to compare the prediction accuracies of CODEAFCS and that of CIBERSORTx for each cell type in each dataset. $\star\star\star$ denotes p-values $<$ 2e-16.

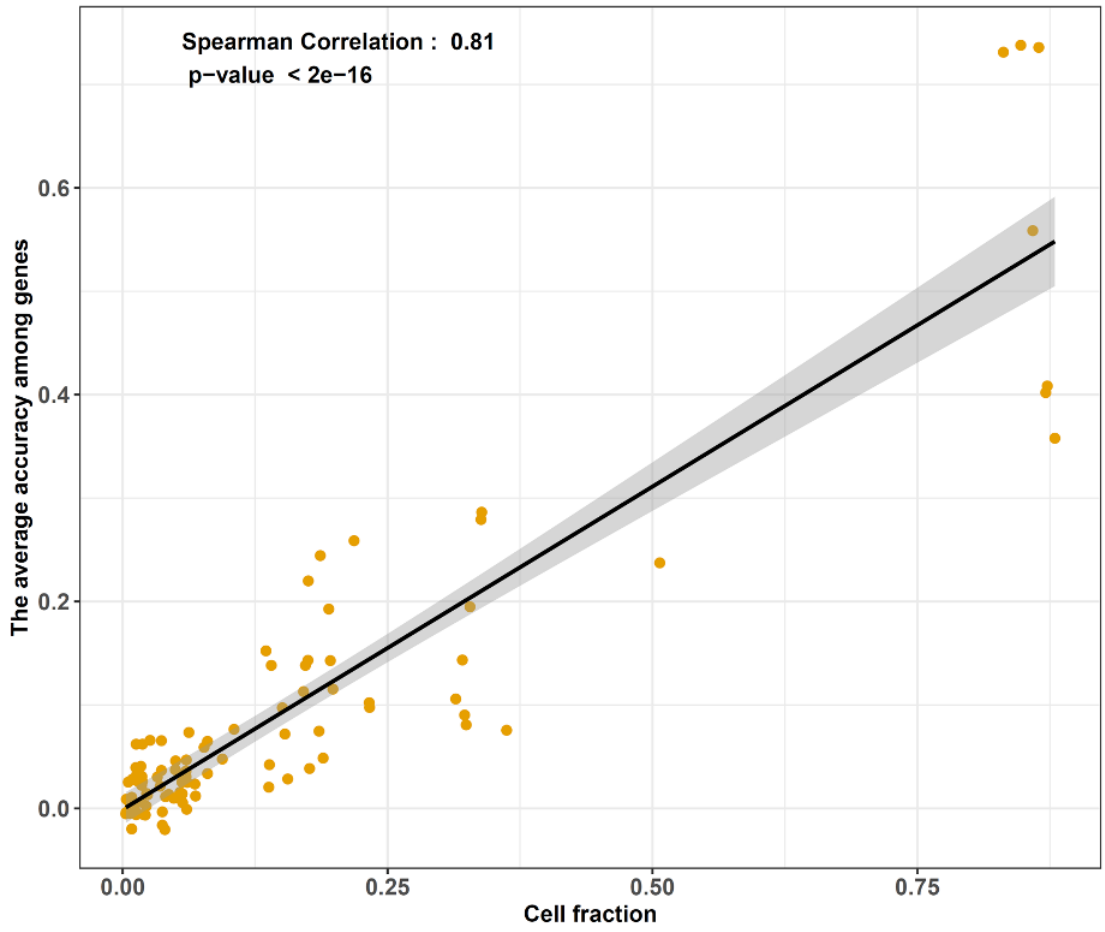

Figure 5.11: For each cell type of each dataset, the average prediction accuracy was computed by taking the average of prediction accuracies (Kendall correlation) across all genes. The y-axis indicates the average prediction accuracy among genes and the x-axis indicates the cell fraction. The Spearman correlation coefficient is 0.81 (p-value $<$ 2e-16).

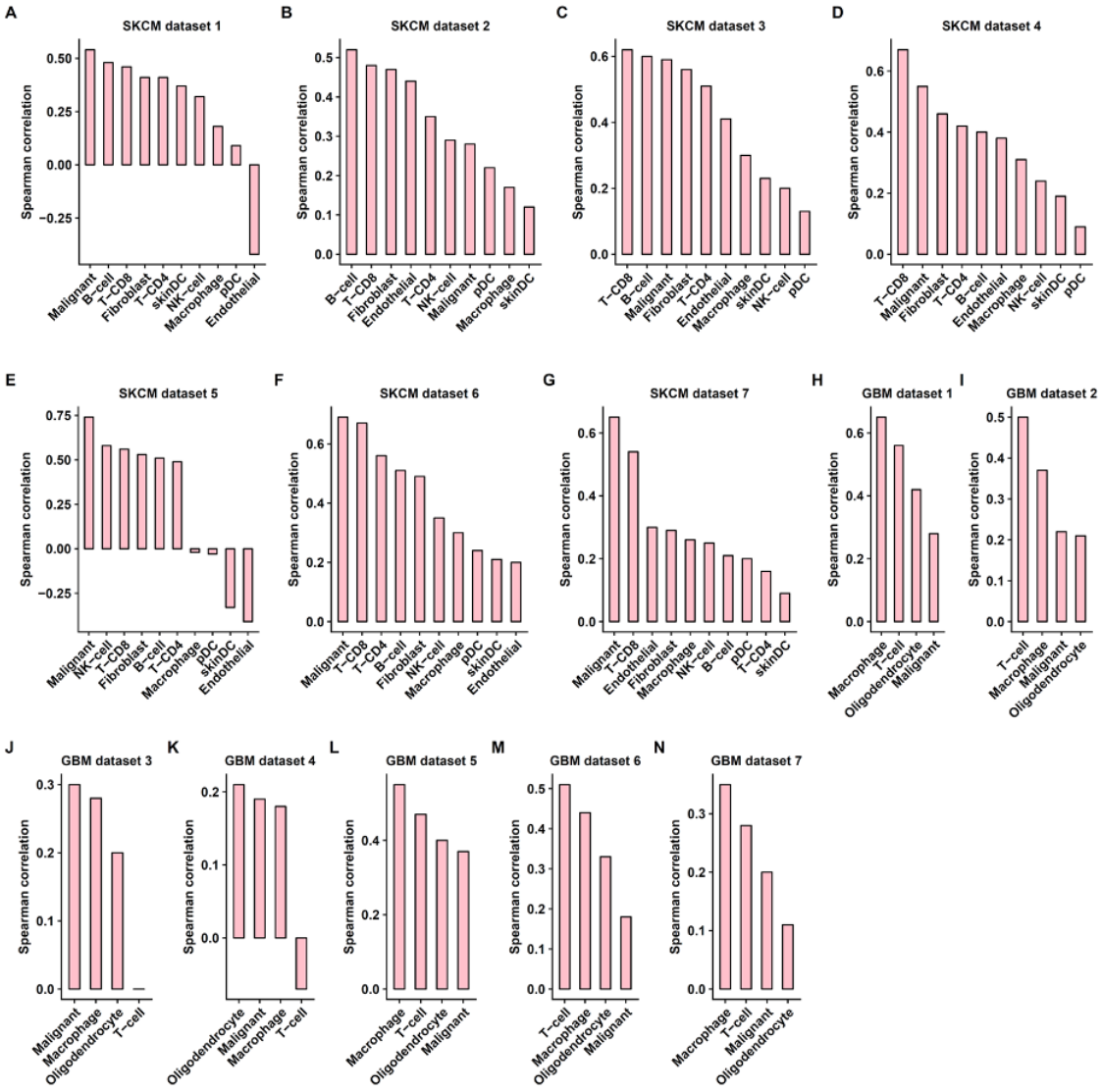

Figure 5.12: (A-N) bar plots depicting the Spearman correlation between prediction accuracies and confidence scores in each cell type for all 14 benchmark datasets (SKCM dataset 1, SKCM dataset 2, SKCM dataset 3, SKCM dataset 4, SKCM dataset 5, SKCM dataset 6, SKCM dataset 7, GBM dataset 1, GBM dataset 2, GBM dataset 3, GBM dataset 4, GBM dataset 5, GBM dataset 6, GBM dataset 7). The y-axis indicates the Spearman correlation value, while the x-axis indicates the cell types.

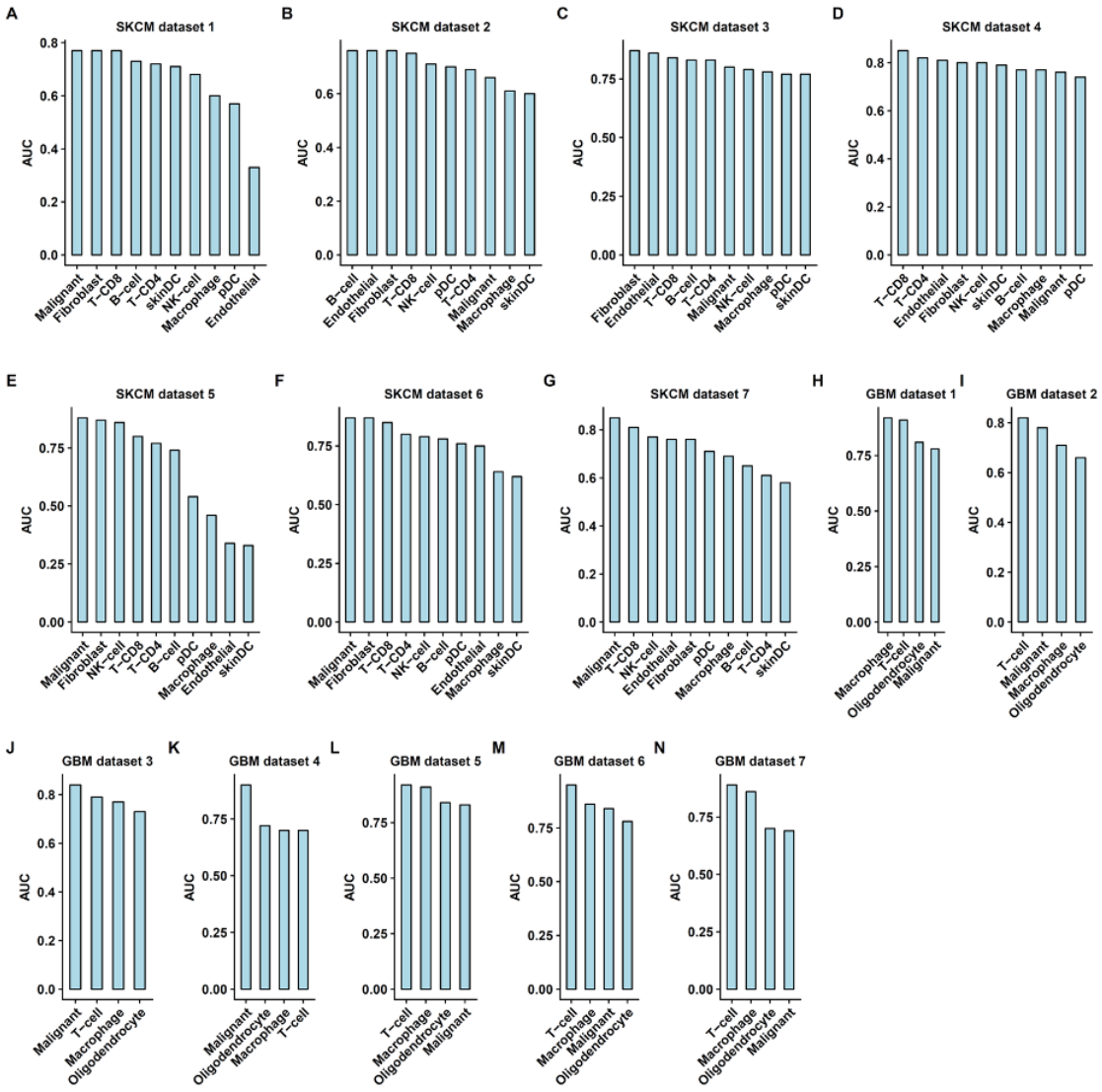

Figure 5.13: (A-N) bar plots depicting the AUC in each cell type for all 14 benchmark datasets (SKCM dataset 1, SKCM dataset 2, SKCM dataset 3, SKCM dataset 4, SKCM dataset 5, SKCM dataset 6, SKCM dataset 7, GBM dataset 1, GBM dataset 2, GBM dataset 3, GBM dataset 4, GBM dataset 5, GBM dataset 6, GBM dataset 7). Genes in each cell-type are grouped into two classes based on the correlation between their predicted and actual expression, informative (prediction accuracy ≥ 0.1 and p-value ≤ 0.05) and uninformative (prediction accuracy < 0.1 or p-value > 0.05). The y-axis indicates the AUC, while the x-axis indicates the cell types.

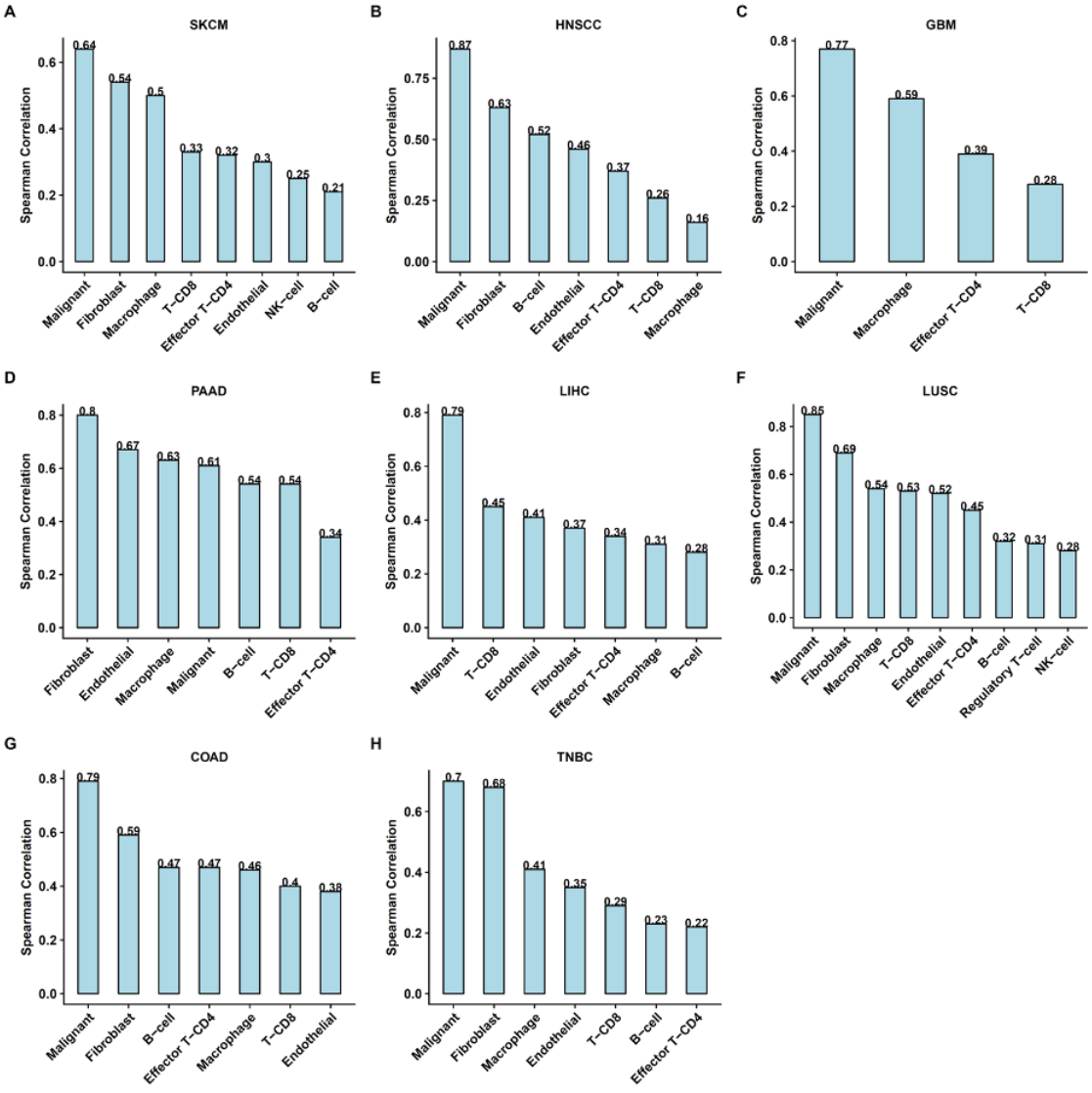

Figure 5.14: (A-H) bar plots depicting the result for SKCM, HNSCC, GBM, PAAD, LIHC, LUSC, COAD and TNBC respectively. The y-axis indicates the Spearman correlation coefficient value, and the x-axis indicates the cell type.

5.3.3 Tumors with DNA mismatch repair deficiency have heightened T-cell co-stimulation that is independent of their tumor mutation burden levels

In normal cells, DNA is constantly repaired in response to DNA damage or DNA replication errors [63]. However, defects in specific DNA repair pathways in cancer cells may result in the accumulation of many somatic mutations resulting in hypermutated tumors (TMB $\geq$ 10-20 mutation/Mb) [133, 4, 160]. One of the sources of hypermutability is a mismatch repair deficiency (MMRD), which leads to the accumulation of insertions and deletion mutations in microsatellite regions of the genome due to uncorrected DNA replication polymerase slippage events. This is known as microsatellite instability (MSI) [67, 52]. Solid tumors with mismatch repair deficiency were shown to be sensitive to immune checkpoint blockade (ICB) therapy irrespective of tumor type, leading the FDA to approve MSI as the first cancer type agnostic biomarker for patients receiving anti-PD1 treatment [28]. The reason behind this general sensitivity to anti-PD1 treatment is not completely understood. Prior work has led to the prevailing hypothesis that elevated tumor mutation burden in mismatch repair deficient tumors leads to more neoantigens, and thus is more likely to activate a host immune response against tumor cells [67, 68, 132, 122]. However, not all tumor types with elevated tumor mutation burden have similar response rates to anti-PD1 [252, 89], and recent studies have revealed that T cells recognize and respond to only a few neoantigens per tumor [193, 20, 197, 108].

More generally, when looking at non-synonymous tumor mutation burden, MSI and survival data across the TCGA collection (Figure **??**, panels A and B, borrowed from [226, 26]. See Methods), we see a significant association between hypermutability and survival benefit of patients in solid tumor types with a frequent underlying mismatch repair deficiency (Figure 5.15, panel C. log-rank test p-value = 0.00084), in contrast to other tumor types (Figure 5.15, panel D, log-rank test p-value = 0.4). These survival differences can be partially explained by the mutual exclusivity of microsatellite instability and chromosomal instability, which has been previously linked with a worse prognosis[226, 56, 13]. Taken together, these findings motivated us to further study cellular immune crosstalk in the tumor microenvironment of mismatch repair deficient tumors to gain additional cell-type-specific insights into their sensitivity to anti-PD1.

We hence aimed to identify cell-cell interactions that are differentially active between microsatellite instable tumors (highlighted as red dots in Figure 5.15, panel A) and microsatellite stable tumors (highlighted as black dots in Figure 5.15, panel A). To this end, we applied CODEFACS to deconvolve the bulk gene expression of all solid tumors from TCGA and integrated their predicted cell-type-specific gene expression levels with LIRICS. The top 50 interactions from this differential analysis (ordered by FDR adjusted p-value) are shown in a network in Figure 5.16 panel A. These interactions are frequently active in mismatch repair deficient tumors of distinct tumor types (Figure 5.17) and importantly, they are more frequently active in hypermutated tumors with DNA mismatch repair deficiency compared to other hypermutated tumors (Figure 5.16, panel B, Figure 5.15, panel A), testifying

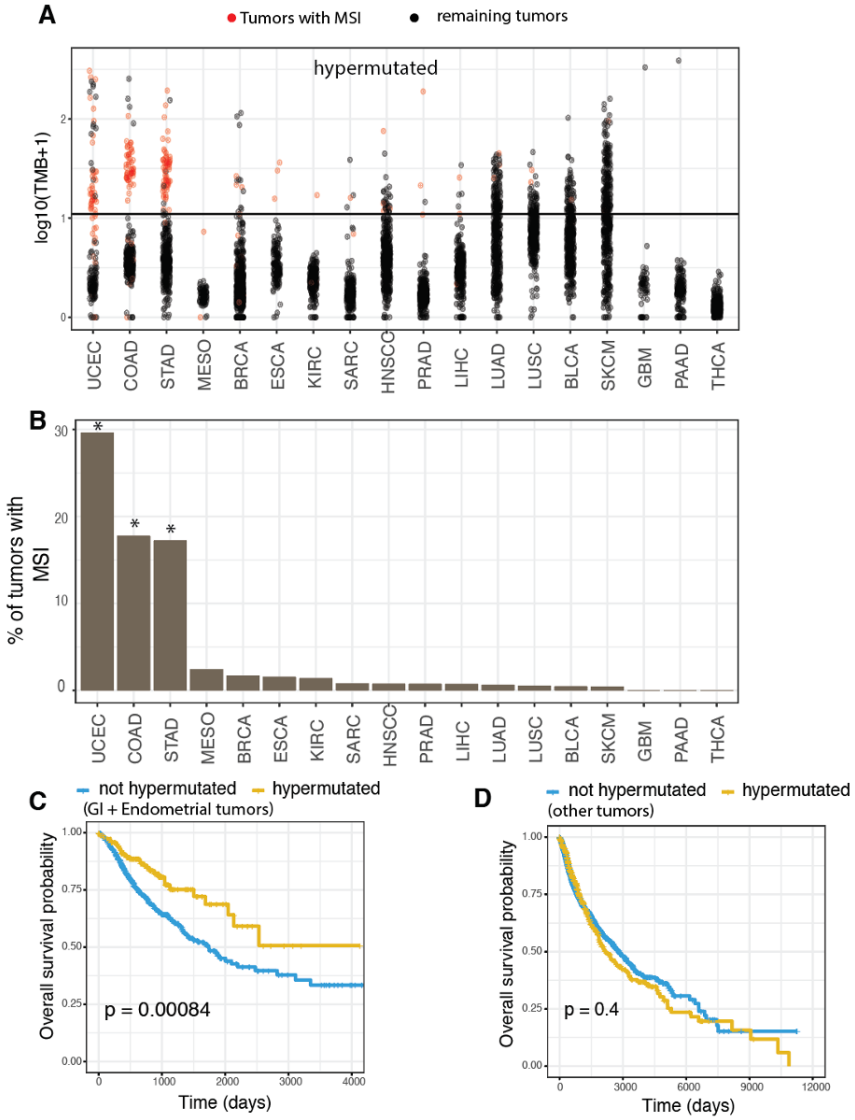

Figure 5.15: (A) This panel plots the distribution of non-synonymous tumor mutation burden on a logarithmic scale (Y-axis). All points above the horizontal line are typically regarded as hyper-mutated tumors (> 10 mutations/Mb). All red points represent tumors with a DNA mismatch repair deficiency detected via microsatellite instability (MSI). (B) This panel depicts the percentage of all tumor samples per cancer type with microsatellite instability (Y-axis). Tumor types marked with a $\star$ represent those where MSI is prevalent. (C-D) Comparison of overall survival of patients with tumors that are hypermutated vs not hypermutated. Left panel (C) In Gastro-Intestinal+Endometrial tumor types where MSI is prevalent (marked with a $\star$ in panel B). Right panel (D) In other solid tumor types where tumors rarely have an underlying mismatch repair deficiency. Statistical significance of differences in survival was calculated using the log-rank test.

to their MSI specificity. The top 50 MSI-specific interactions include the PDL1-PD1 checkpoint interaction between tumor cells and CD8+ T-cells, but notably, T-cell activating/co-stimulatory interactions such as the 41BBL-41BB interaction between Tumor cells and CD8+ T cells, ULBP2-NKG2D between tumor cells and CD4+T cells, and chemotaxis interactions involved in trafficking of lymphocytes in and around the tumor mass, such as the CXCL9-CXCR3 chemokine interaction between macrophages and CD4+ T cells and CCL3/4/5 – CCR5 interactions between various immune and stromal cell-types.

This shared heightened cellular crosstalk unique to the TME of mismatch repair deficient tumors suggests that tumor infiltrating T cells can be activated by co-stimulatory signals in the TME independent of overall tumor mutation burden, only to be kept in balance by other immunoregulatory mechanisms such as the PD1-PDL1 checkpoint interaction between CD8+ T cells and tumor cells. Our results indicate that when this interaction is blocked by anti-PD1 treatment, the presence of other co-stimulatory interactions can lead to the observed enhanced response of MMRD tumors to immune checkpoint blockade therapy. This in turn raises the possibility that switching on specific T cell co-stimulation signals in the TME may lead to better responses to anti-PD1 treatment independent of tumor mutation burden. Notably, recent pre-clinical studies have shown that combination therapies aimed at enhancing such T-cell co-stimulating interactions improve anti-tumor immune responses even in low TMB and highly immuno-suppressive settings [42, 135, 18, 147, 43, 177]. Currently, several clinical trials to assess the safety and efficacy of these combinations are in progress [215, 229, 51, 46].

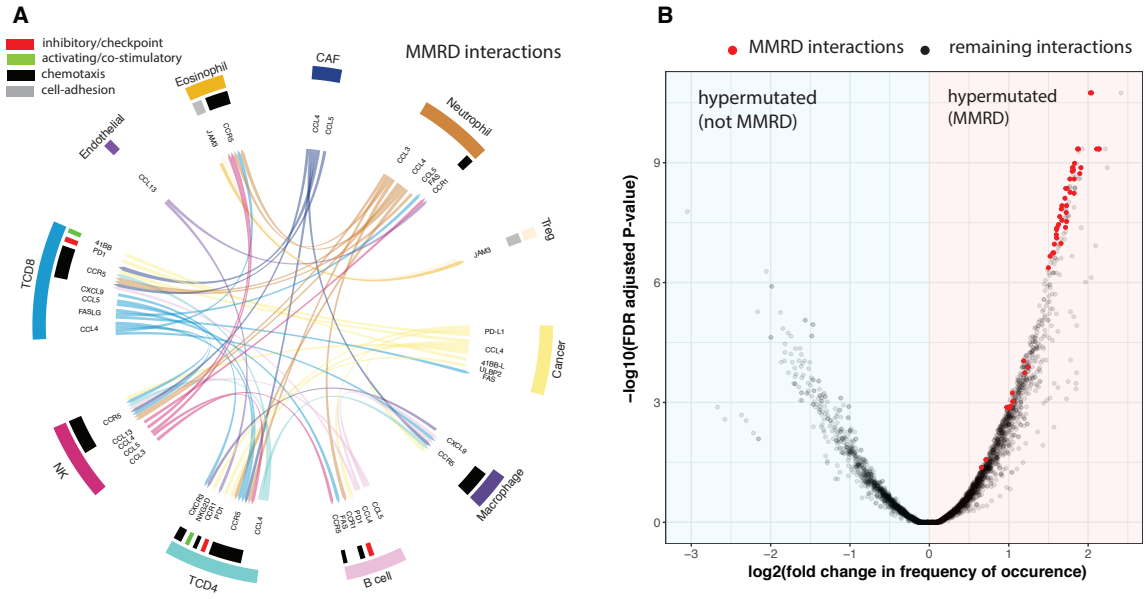

Figure 5.16: (A) Interaction network consisting of the top 50 interactions. Interactions highlighted in green represent co-stimulatory interactions/having an activating effect on the target cell. Interactions highlighted in red represent checkpoint interactions/having an inhibitory effect on the target cell. Interactions highlighted in black represent pro-inflammatory/chemotaxis interactions involved in inflammatory response and immune cell trafficking to tumor sites. Eos: Eosinophils, CAF: Cancer associated fibroblasts. (B) A volcano plot depicting on the x-axis the $\log_2$ fold change in the frequency of occurrence of each cell-cell interaction in the TME of hypermutated tumors with an underlying DNA mismatch repair deficiency vs other hypermutated tumors. The y-axis indicates the -log10 FDR adjusted p-value of the observed enrichment. Highlighted in red in the scatter plot are the top 50 interactions that are most differentially active between all MSI vs non-MSI tumors (shown in panel A)

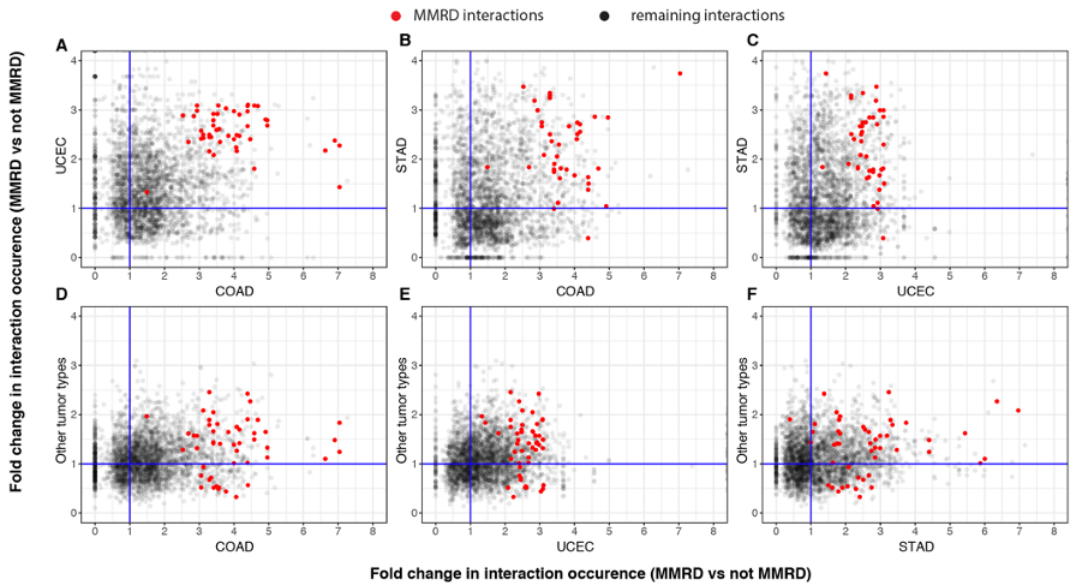

Figure 5.17: (A-F) Tumors are grouped into four different groups by tissue of origin: STAD (stomach), COAD (colon), UCEC (endometrium) and other (all other solid tumor types) in order to have sufficient numbers of mismatch repair deficient vs mismatch repair proficient samples per group. The axes measure the enrichment scores of all plausible ligand-receptor interactions between cell types in their respective group. A fold change > 1 implies the interaction occurs more frequently in mismatch repair deficient tumors. Interactions highlighted in red represent the shared core set of interactions from Figure 5.16 A that are universally enriched in mismatch repair deficient solid tumors.

5.3.4 Machine learning guided discovery of cellular crosstalk predictive of response to immune checkpoint blockade therapy

Given the shortage of large publicly available transcriptomics datasets of patients receiving immune checkpoint blockade therapy, we asked if we can effectively utilize this large resource of deconvolved TCGA data we generated to transfer-learn cell-cell interactions robustly predictive of response to immune checkpoint blockade therapy. Specifically, since some mutations during tumor evolution can be immunogenic, we hypothesized that one could potentially discover cell-type-specific ligand-receptor interactions predictive of response to ICB therapy by a joint analysis of mutation and deconvolved expression data from the TCGA. We focus on

melanoma, currently the tumor type best responding to ICB, where there are many independent publicly available bulk expression datasets of patient's receiving anti-PD1 treatment and single-cell derived cell-type-specific signatures, which serve as priors for the deconvolution of these bulk datasets. Starting from the deconvolved TCGA-SKCM dataset as our training set (N=469), we employed a genetic algorithm to find cell-type specific ligand-receptor interactions, whose activation state best separates hypermutated melanoma tumors from non-hypermutated tumors (Fig 5.15, panel A), assuming that the former group captures more samples with some immunogenic mutations than the latter (Figure 5.18, panel A). We term the interactions identified in this process melanoma mutation specific functional interactions (MSFI), and the network formed by these interactions is displayed in Figure 5.18 panel B.

Having identified the MSFI interactions, we applied CODEFACS to deconvolve the bulk expression data of pre-treatment samples from the three largest publicly available melanoma datasets where patients received anti-PD1 treatment (either monotherapy or in combination with anti-CTLA4; Methods) 42–44. We then employed LIRICS to the respective deconvolved expression of each of these checkpoint datasets, without any additional training, and simply quantified the number of MSFI interactions that are active in each of these patients' tumor samples, which we denote as the tumor's MSFI score. Remarkably, we find that the MSFI score of each sample can robustly stratify patients into those that are likely to respond to ICB vs those that are unlikely to respond (Figure 5.18, panel C, progression free survival log rank test p-value: 0.00057, Figure 5.18, panel D, overall survival log

rank test p-value: 0.0031). Figure 5.19 depicts the survival differences for the two treatment groups separately (anti-PD1 monotherapy and anti-CTLA4 + anti-PD1 combination). Additionally, Figures 5.20 and 5.21 depict the survival differences for each ICB dataset separately. As evident, our results improve over recent bulk gene expression based predictors of melanoma ICB therapy response (IMPRES [14], TIDE [105] and the melanocytic plasticity signature (MPS) scores [176]). We note that the performance levels of the latter on bulk expression datasets, where their original RNAseq reads have been uniformly aligned and normalized as described in the Methods, is lower than that reported in the original publications, pointing to the potential sensitivity of expression-based predictors to the processing used and the need to do that in a uniform, generally accepted manner (see Discussion).

To further evaluate the predictive performance of the MSFI score, we tested its ability to predict partial or complete responders vs stable or progressive disease patients in these datasets. To this end we plotted the receiver operator area under the curve (AUC) obtained using it for classifying the patients to partial or complete responders vs stable or progressive disease, and compared its performance to that obtained with the three other published predictors for the different treatment groups (Figure 5.18, panel E). On average, the MSFI score achieves an AUC of 0.63 (for anti-PD1 monotherapy the AUCs obtained are 0.6, 0.77 and 0.52 for the three individual ICB datasets, and for the combination ICB treatment its 0.77, 0.49 and 0.63). A similar performance could not be achieved if the placement of the ligand and receptor between interacting cell-types in the MSFI network was swapped (average AUC 0.58) or by randomly shuffling the interaction activity profiles (average AUC $\approx$

0.5), testifying that the selected cell-cell interactions are best predictive of response to ICB. Comparing MSFI predictive performance in this response classification task to that of the recent melanoma bulk expression-based predictors, TIDE achieves an average AUC of 0.6 (for anti-PD1 monotherapy the AUCS are 0.46, 0.66 and 0.45 for the three ICB datasets and 0.89, 0.55 and 0.61 for the combination), IMPRES achieves an average AUC of 0.55 (for anti-PD1 monotherapy the AUCS are 0.59, 0.64 and 0.6 respectively for the three ICB datasets and 0.61, 0.44 and 0.42 for the combination) and MPS achieves an average AUC of 0.51 (for anti-PD1 monotherapy the AUCS are 0.61, 0.49 and 0.63 respectively for the three ICB datasets and 0.41, 0.59 and 0.38 for the combination).

Examining the MSFI network leads to interesting insights (Figure 5.18, panel B). First, we find an over-representation of cell-type-specific co-stimulatory/immune cell activating interactions known from prior immunological literature (hypergeometric test p-value < 0.05) [23, 219, 159, 41, 35, 174, 242, 80, 220, 40, 235]. Second, the MSFI network additionally includes cytokine/chemokine interactions involved in pro-inflammatory response and the trafficking of NK, T and B cells to the tumor site (responsible for better lymphocyte infiltration into the tumor mass). Importantly, on fitting a multivariate Cox-proportional hazards model with MSFI scores and TMB of each patient receiving anti-PD1 (wherever TMB data was available), we see that a high MSFI score is significantly associated with improved progression free survival (p-value: 0.013) and overall survival (p-value: 0.0258), whereas high TMB is not (PFS p-value: 0.224, OS p-value: 0.477). These results further support the findings from the TCGA analysis that heightened co-stimulatory and

pro-inflammatory signals in the TME can mobilize tumor infiltrating lymphocytes to generate an effective host immune response upon immune checkpoint blockade independent of TMB.

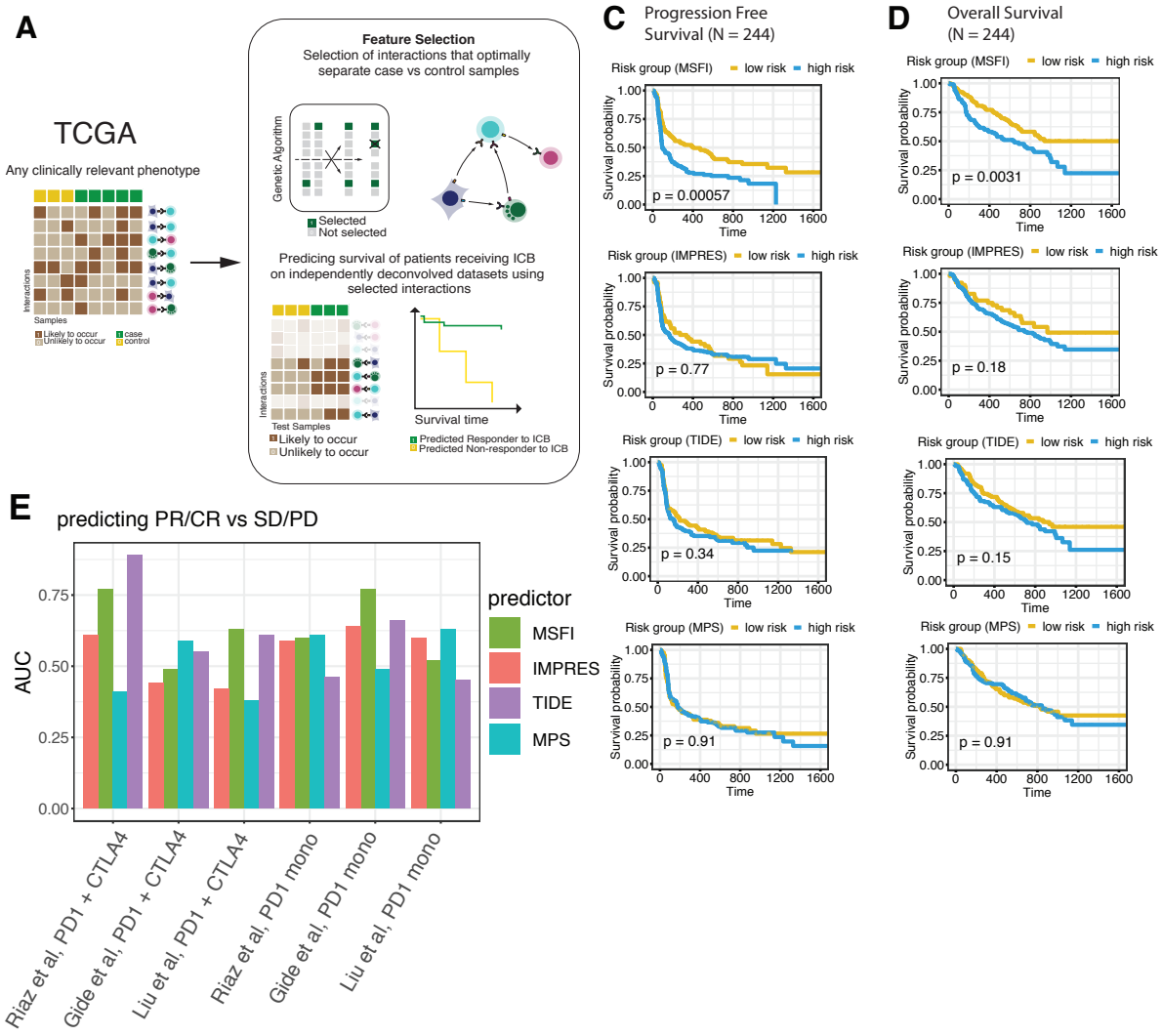

Figure 5.18: (Continued on next page)

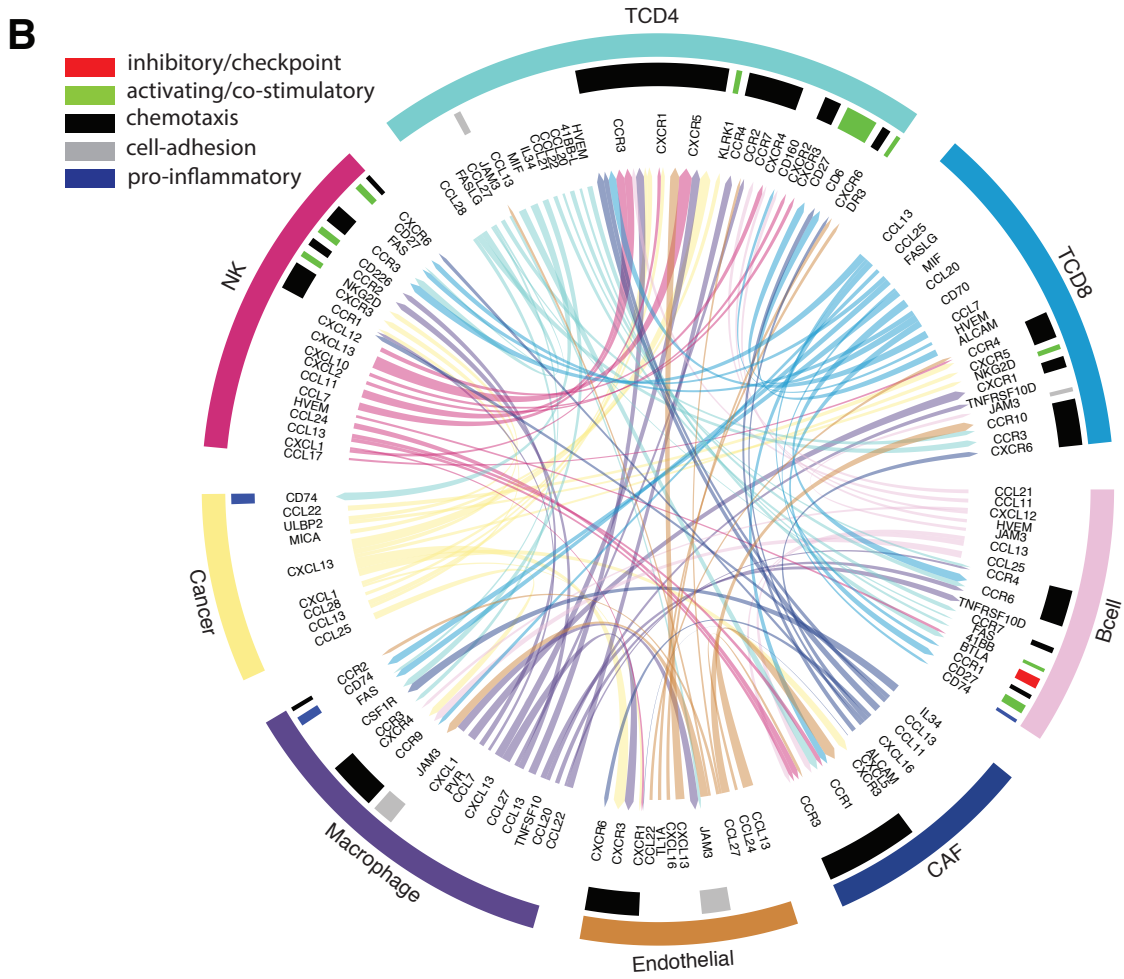

Figure 5.18: (Continued on next page)

Figure 5.18: (A) Overview of the machine learning analysis employed to identify cell type specific interactions that are predictive of response to immune checkpoint blockade therapy. (B) A chord diagram of the resulting MSFI network. Each individual interaction is represented by a link from the source cell type (ligand expressing cell type) to the target cell type (receptor expressing cell type) and the color of the link represents the color of the source cell type. For interactions that are activating/co-stimulatory, the sector in the corresponding target cell type is highlighted in green. For inhibitory/checkpoint interactions, the sector in the target cell type is in highlighted red. Interactions involved in chemotaxis are highlighted in black and those I mediating a pro-inflammatory response are highlighted in blue, cell-adhesion interactions are highlighted in grey. (C) Kaplan-Meier plot depicting the progression free survival of the combined set of melanoma patients receiving immune checkpoint blockade (N= 244). On the top, the patients are stratified into low-risk/high-risk groups based on the median value of MSFI score from LIRICS. Second from top, patients stratified into low/high risk groups based on median IMPRES score[14]. Third from top, patients stratified into low/high risk groups based on median TIDE score[105], Bottom, patients stratified into low/high risk groups based on median MPS score [176]. (D) Kaplan-Meier plots depicting the overall survival of all melanoma patients receiving immune checkpoint blockade (N= 244). On the top, the patients are stratified into low-risk/high-risk groups based on the median value of MSFI score from LIRICS. Second from top, patients stratified into low/high risk groups based on median IMPRES score[14]. Third from top, patients stratified into low/high risk groups based on median TIDE score[105], Bottom, patients stratified into low/high risk groups based on median MPS score [176]. Survival differences among patients that received anti-PD1 monotherapy vs anti-CTLA4 + anti-PD1 combination are shown in supplementary figure 10 (E) Area under the ROC curves in predicting Complete/Partial-response (based on RECIST v1.1) to immune checkpoint blockade therapy for the different scores. X-axis marks patients grouped by dataset source and treatment regimen. PD1 mono represents patients that received anti-PD1 monotherapy. PD1 + CTLA4 represents patients that additionally received anti-CTLA4 besides anti-PD1.

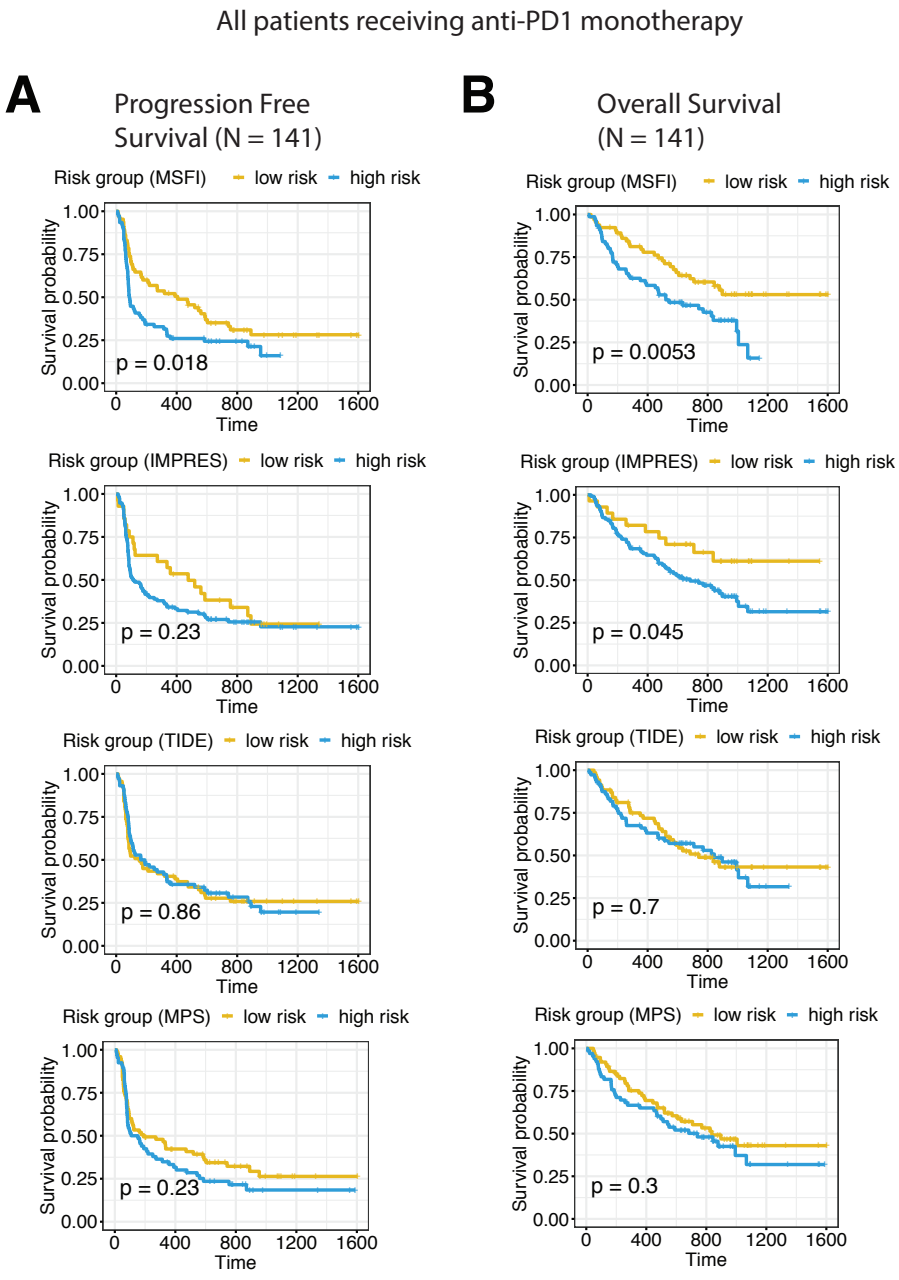

Figure 5.19: (Continued on next page)

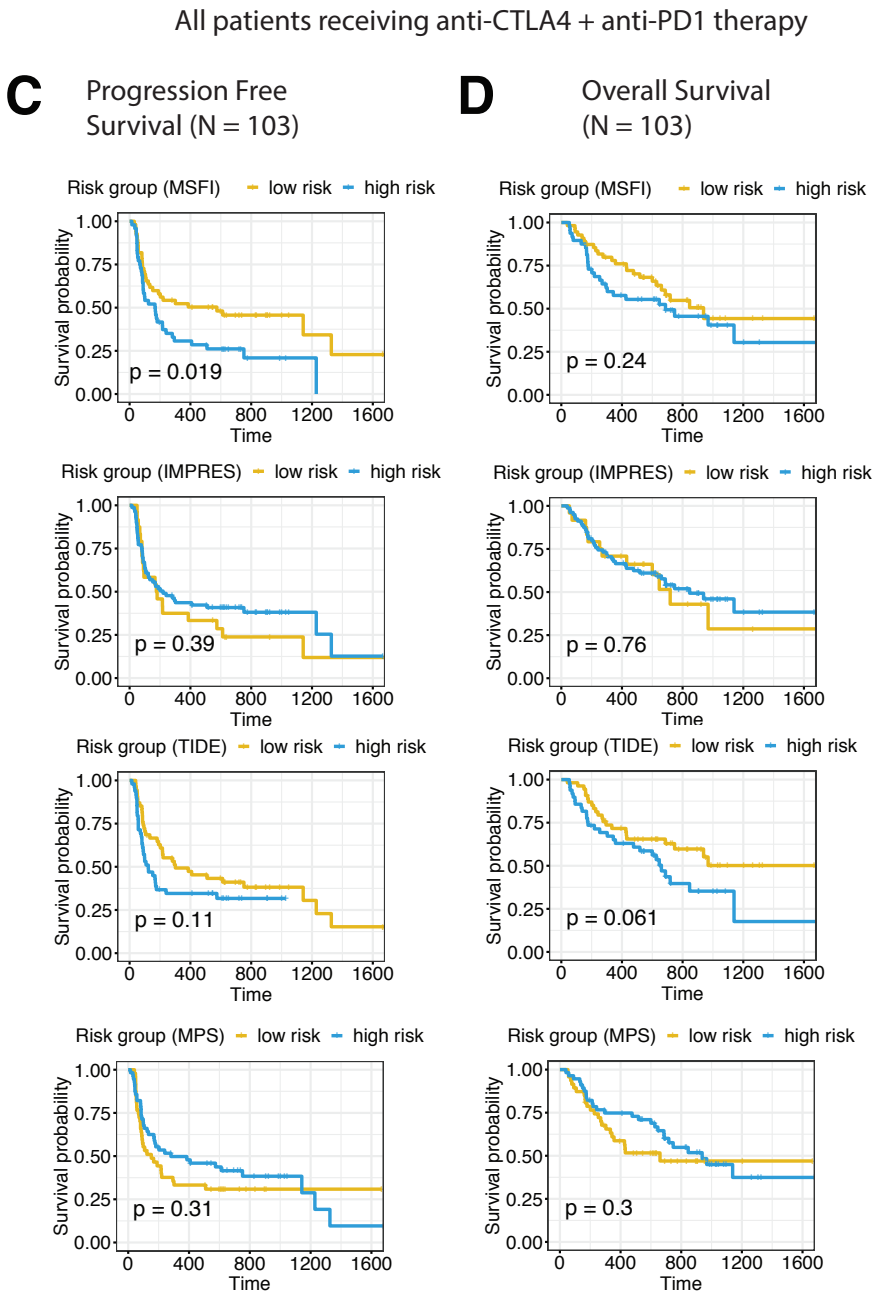

Figure 5.19: All patients receiving anti-PD1 treatment are classified into low-risk group if their LIRICS based cellular crosstalk score (MSFI score) exceeds the population median. The Likewise, when using the IMPRES score [14]. For TIDE [105] and MPS scores [176], all patients receiving anti-PD1 treatment are classified into low-risk group if their values fall below the population median (as these scores were shown to be associated with immune resistance1,2) (A,B) depict the survival differences for patients receiving anti-PD1 monotherapy only. (C,D) depict the survival differences for patients receiving anti-CTLA4 + anti-PD1 combination. Significance in the difference of survival trends of the two groups is calculated using the log-rank test. Time on the x-axis is measured in days.

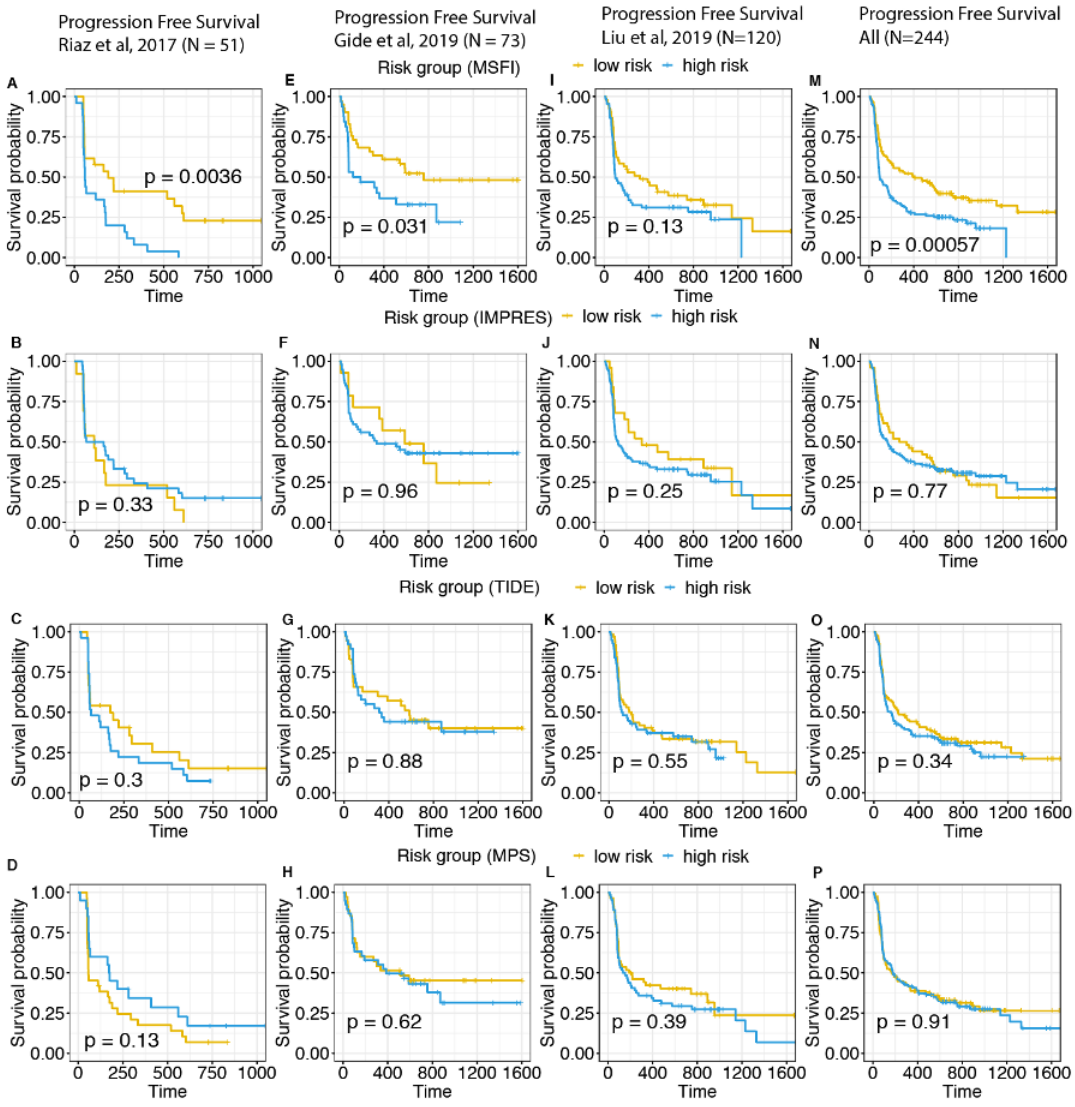

Figure 5.20: (A, E, I, M) Patients are classified into low-risk group if their LIRICS based cellular crosstalk score exceeds the population median. (B, F, J, N) Patients are classified into low-risk group if their IMPRES score [14] exceeds the population median. (C, G, K, O) Patients are classified into low-risk group if their TIDE score [105] is less than the population median. (D, H, L, P) Patients are classified into low-risk group if their MPS score [176] is less than the population median. Significance in the difference of survival trends of the two groups is calculated using the log-rank test. Time on the x-axis is measured in days.

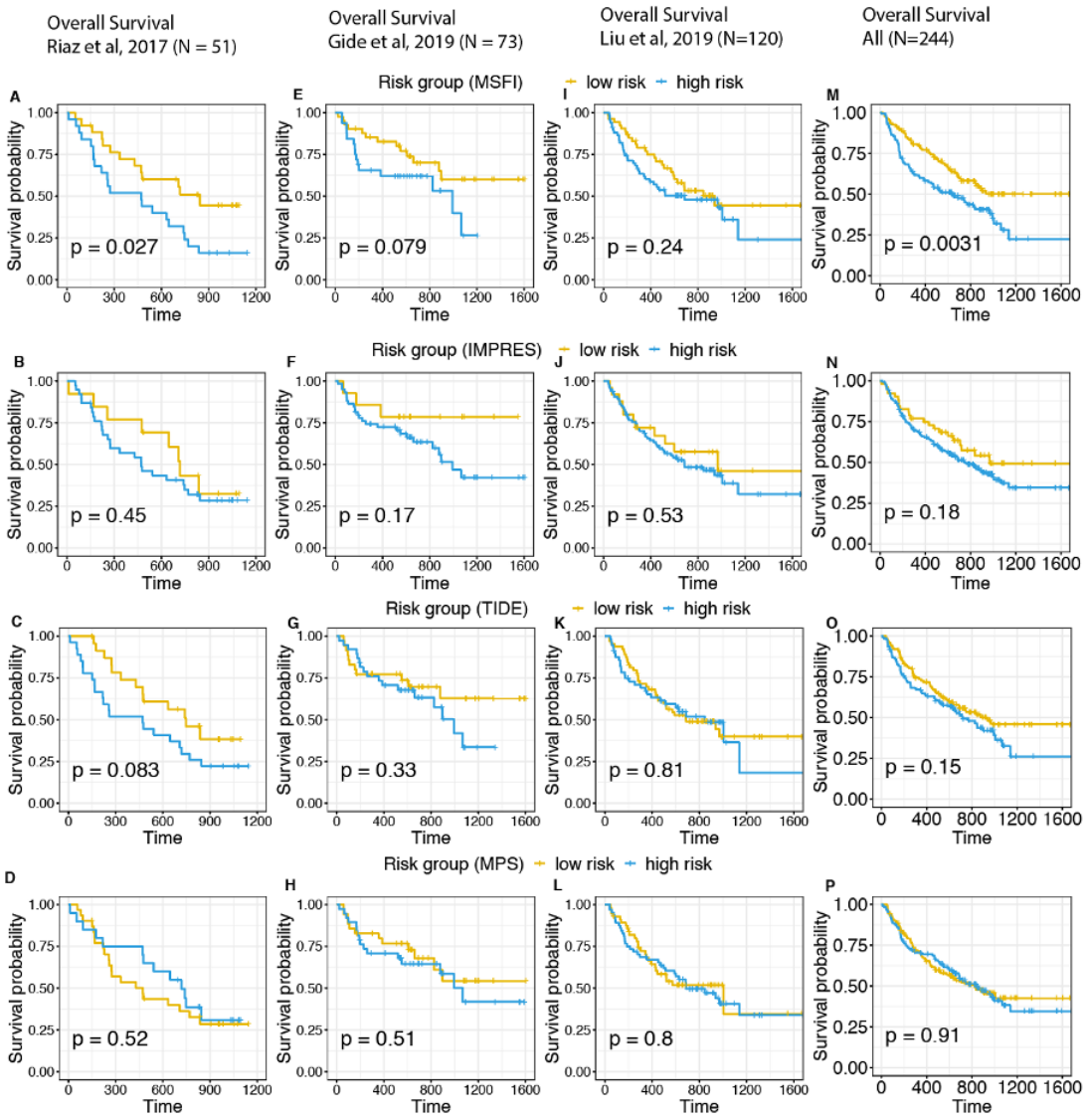

Figure 5.21: (A, E, I, M) Patients are classified into low-risk group if their LIRICS based cellular crosstalk score exceeds the population median. (B, F, J, N) Patients are classified into low-risk group if their IMPRES score [14] exceeds the population median. (C, G, K, O) Patients are classified into low-risk group if their TIDE score [105] is less than the population median. (D, H, L, P) Patients are classified into low-risk group if their MPS score [176] is less than the population median. Significance in the difference of survival trends of the two groups is calculated using the log-rank test. Time on x-axis is measured in days.

5.4 Discussion

This study presents a computational tool, CODEFACS and a pipeline, LIRICS, that enables an (averaged) 'virtual single cell' characterization of the TME from bulk tumor expression data. Applying these tools, we identify cell type specific ligand-receptors interactions that are active and functionally important within individual tumor microenvironments, in modifying patients' survival and response to ICB. Applying CODEFACS to 8000 tumors from TCGA, we estimate the cell-type-specific gene expression profiles of each individual tumor sample thus enabling the analysis of the TCGA at a cell-type-specific resolution. Integrating these data with LIRICS, we systematically characterized the immune cellular crosstalk of the tumor microenvironments of different tumor types. We identified a shared core of intercellular TME interactions in DNA mismatch repair deficient tumors, which are associated with improved patient survival and high sensitivity to immune checkpoint blockade therapy. One potentially interesting implication of these findings is that immunomodulators enhancing T-cell co-stimulation (e.g, via the 41BB receptor) might improve patients' response to ICB irrespective of their tumor mutation burden. Finally, focusing on melanoma, we show that one can bootstrap on the large deconvolved data resource from TCGA using machine learning techniques to discover cell-cell interactions within the TME that successfully predict patients' response to immune checkpoint blockade.

Now, while we have provided a toolkit to discover clinically relevant cell-cell interactions from bulk tumor expression, there are some important limitations that

should be noted and potentially further improved upon in the future. First, it requires prior information about the cell type composition of the input tumors, or alternatively, knowledge of the pertaining cell-types' gene expression or methylation signatures that can be used to infer their abundances, and its accuracy depends on the accuracy of the latter. Second, its prediction power is limited to subsets of the whole exome genes, and its performance deteriorates for lowly-abundant cell types. However, the confidence scores provided partially alleviate this limitation, allowing the user to rank genes in each cell-type by the quality of predictions of the expression of each gene in a given cell-type. Third, regarding LIRICS, it is currently restricted to well defined immune related ligand-receptor interactions between tumor, immune, stromal and epithelial cell types and does not consider the spatial localization of cells in the TME. The inclusion of the latter with the advent of forthcoming spatial transcriptomics data is likely to lead to considerably more informative interaction inference approaches.

(The source code related to this work is currently under review for a patent. All codes and data will be made publicly available upon publication.)

Chapter 6

Conclusions

Tumor heterogeneity is a significant hurdle to developing cures for cancer. Hence, to keep up with this constantly evolving enemy, it is important to develop computational methods that parse these high dimensional genomic datasets and unearth underlying patterns that can help us make sense of this heterogeneity. In this work, we present some concrete examples.

6.0.1 Contributions to our understanding of epigenetic heterogeneity in cancer

In chapter 2, we developed a computational framework to explore how epigenetic heterogeneity in cancer cells may lead to functional rewiring of genes via dynamic changes in the protein-protein interaction network. With the help of breast cancer and adjacent normal breast tissue gene expression data from TCGA, we show how functional re-wiring events that are frequently selected in cancer modulate patient survival in general and lead to an improved clustering of clinically relevant breast cancer subtypes. Such functional heterogeneity explains why many genetic interactions are context specific, making their translation to clinically effective targeted therapies a challenge. Currently, the exception to these trends are synthetic lethal interactions between single strand and double strand break DNA repair path-

ways, which are highly conserved in all eukaryotic cells [11]. However, there has been renewed interest in the identification of additional robust genetic interactions in tumor cells with the emergence of high throughput CRISPR screening technologies [59, 78, 10, 183, 165, 210]. Another aspect fundamental to characterization of epigenetic heterogeneity is our ability to model how cells respond to different genetic or environmental perturbations. A pre-requisite to building such models is having good functional annotations of biological networks. However, such annotations are currently sparse [213]. To make some progress in this direction, we developed and validated new mixed integer linear programming formulations that utilize high-throughput genetic screening data and the network topology to annotate biological networks of cells with directions of signal flow and signs representing different functional activation or inhibitory effects. Overall, we demonstrated that our method markedly outperforms the state of the art for this task.

6.0.2 Contributions to our understanding of genetic heterogeneity in cancer

In chapter 3 we shifted gears to understand how intra-tumor genetic heterogeneity impacts anti-tumor host immune responses in melanoma. From the computational side, we developed an unbiased approach to estimate the number of genetically distinct cancer cell clones in any given tumor sample by utilising both point mutation and copy number alteration information on each sample. Our results from patient and mice data suggest that increased intra-tumor genetic heterogene-

ity leads to reduced overall immunogenicity of the tumor sample despite similar or higher levels of tumor mutation burden thereby impairing host anti-tumor immune responses. These findings are clinically relevant as the FDA recently approved tumor mutation burden as a biomarker for responses to anti-PD1 treatment in metastatic solid tumors [224]. Our results indicate that an elevated mutation burden may lead to poorer responses to immune checkpoint blockade therapies in melanoma if accompanied by an increased intra-tumor genetic heterogeneity. Hence, moving ahead, it is going to be important to account for the intra-tumor genetic heterogeneity of the sample when using the tumor mutation burden as a biomarker to decide which patient should receive immune checkpoint blockade treatments. In addition, in chapter 4, we came up with statistical machine learning techniques to uncover novel factors explaining the observed heterogeneity of recurrent chromosome arm gains and losses in cancer. Overall, our analysis of the GTEx and TCGA databases revealed that the normal transcriptional state of different chromosome arms in a tissue can influence which arm is recurrently gained or lost in emerging cancer types from that tissue.

6.0.3 Contributions to our understanding of micro-environmental heterogeneity in cancer

In chapter 5, we developed a new computational tool CODEFACS that markedly improves over the state of the art method, CIBERSORTx, in reconstruction of cell type specific transcriptomes from bulk gene expression profiles of each tumor sam-

ple. With this tool we are not only able to infer the abundance of different cell types in each sample, but also the transcriptional states they exist in. This information is key for deciphering which cell types are likely to interact with each other in each patient's tumor micro-environment and which of these interactions are likely to modulate responses to immunotherapy. Using our tool we deconvolved the TCGA collection to provide a cell type specific molecular atlas of $\approx$ 8000 tumor samples. This resource can serve as a valuable test bed for the immuno-oncological research community to test specific hypotheses about specific cell types and the effect of their interactions with other cell types on clinical outcomes of patients.

Overall, our analysis of the TCGA and immune checkpoint blockade treated datasets using CODEFACS + LIRICS reveals that while tumor neo-antigens are a necessary "ignition switch" to activate T cells, additional co-stimulatory signals in the TME (for eg: 41BB-41BBL, ULBP2-NKG2D) might be required to sustain effective anti-tumor immune responses upon releasing the brakes using immune checkpoint blockade treatments. In the future, finding the right balance is going to be key to maximizing clinical benefit of patients while minimising immune related adverse events.

Although this work focuses on studying the tumor microenvironment, these methods can be applied to discover important cell-cell interactions in noncancerous tissues under a variety of normal and disease states. One interesting application that we envision is the characterization of clinically relevant intercellular interactions occurring at the maternal-fetal interface using corresponding bulk gene expression data and pregnancy outcome information, whose elucidation may help treat and

mitigate preeclampsia and other pregnancy related complications. One can also use our tools to study bulk gene expression data from pre-malignant tissue samples and compare them against malignant samples to elucidate cell-cell interaction dynamics on the way malignancy. Finally, one can use our tools to deconvolve expression data from autoimmune disorders to learn more about the underlying immune interactions.

6.0.4 The challenges and road ahead

The rate at which genomic datasets are being generated is gradually increasing. Furthermore, with the emergence of self-supervised deep learning [106], in principle we should be able to directly extract features associated with specific patterns of genomic alterations in patients and overlay them on top of clinical data to identify new biomarkers. However, there are two fundamental challenges unique to this domain that need to be addressed. First, is the lack of standardized data pre-processing methodologies and second, is the curse of high dimensionality, which can lead to model overfitting. For instance, several recent publications have reported discrepancies between different RNA-seq expression quantification methods based on the sample preservation protocol, reference transcriptome version used and choice of method (alignment based vs alignment free) [250, 203, 64, 227, 194]. This can dramatically affect reproducibility of the learned predictive models. While in fields like computer vision and natural language processing, such issues are addressed by feeding the models with more data from other related domains or data augmentation, such a strategy is still not scalable to genomics datasets given their incredible high

dimensionality. To address this challenge, injecting prior mechanistic or expert knowledge into the feature representations would be key [12].

With the emergence of methods that integrate genomic knowledge with non-invasive imaging or liquid biopsy technologies [47, 141, 94], accessing specific features of each patient's tumor and the surrounding micro-environment will only get easier. This will enable real-time monitoring of patient responses from diverse perspectives and hence provide a much more pragmatic and scalable framework for guiding precision medicine-based treatment combinations.

Although development of methods discussed above will make management of tumor heterogeneity easier in clinics, it will still not solve the problem. Additional effort should also be invested in basic science; understanding the mechanisms fueling tumor heterogeneity. For example, it is important to study the mechanisms of generation and re-integration of extrachromosomal DNA elements into cancer genomes, which is a major driver of intra-tumor genetic heterogeneity and treatment resistance [114]. This could reveal novel ways to uniquely target cancer cells so that they don't diversify and bounce back in response to standard of care treatments.

Bibliography

[1] The Cancer Genome Atlas, 2020.

[2] J. Ahn, Y. Yuan, and G. Parmigiani. Demix: Deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871.

[3] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, Milind B. Suraokar, Lixia Diao, Ignacio I. Wistuba, and Wenyi Wang. DeMix: Deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871, 2013.

[4] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, G.R. Bignell, N. Bolli, A. Borg, and A.L. Borresen-Dale. Signatures of mutational processes in human cancer. *Nature*, 500:415–421.

[5] U. Alon. Biological networks: The tinkerer as an engineer. *Science*, 301(5641):1866–1867, 2003.

[6] G. Alter, J.M. Malenfant, and M. Altfeld. Cd107a as a functional marker for the identification of natural killer cell activity. *Journal of immunological methods*, 294:15–22.

[7] N. Andor, T.A. Graham, M. Jansen, L.C. Xia, C.A. Aktipis, C. Petritsch, H.P. Ji, and C.C. Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine*, 22:105–113.

[8] D. Aran. Cell-type enrichment analysis of bulk transcriptomes using xcell. *Methods in Molecular Biology*, 2020:263–276.

[9] Dvir Aran, Roman Camarda, Justin Odegaard, Hyojung Paik, Boris Oskotsky, Gregor Krings, Andrei Goga, Marina Sirota, and Atul J Butte. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature communications*, 8(1):1–14, 2017.

[10] Michael Aregger, Keith A. Lawson, Maximillian Billmann, Michael Costanzo, Amy H.Y. Tong, Katherine Chan, Mahfuzur Rahman, Kevin R. Brown, Catherine Ross, Matej Usaj, Lucy Nedyalkova, Olga Sizova, Andrea Habsid, Judy Pawling, Zhen Yuan Lin, Hala Abdouni, Cassandra J. Wong, Alexander Weiss, Patricia Mero, James W. Dennis, Anne Claude Gingras, Chad L. Myers, Brenda J. Andrews, Charles Boone, and Jason Moffat. Systematic mapping of genetic interactions for de novo fatty acid synthesis identifies C12orf49 as a regulator of lipid metabolism. *Nature Metabolism*, 2(6):499–513, 2020.

[11] Alan Ashworth and Christopher J Lord. Synthetic lethal therapies for cancer: what's next after parp inhibitors? *Nature reviews Clinical oncology*, 15(9):564–576, 2018.

[12] Noam Auslander, Ayal B. Gussow, and Eugene V. Koonin. Incorporating machine learning into established bioinformatics frameworks. *International Journal of Molecular Sciences*, 22(6), 2021.

[13] Noam Auslander, Yuri I. Wolf, and Eugene V. Koonin. Interplay between DNA damage repair and apoptosis shapes cancer evolution through aneuploidy and microsatellite instability. *Nature Communications*, 11(1), 2020.

[14] Noam Auslander, Gao Zhang, Joo Sang Lee, Dennie T. Frederick, Benchun Miao, Tabea Moll, Tian Tian, Zhi Wei, Sanna Madan, Ryan J. Sullivan, Genevieve Boland, Keith Flaherty, Meenhard Herlyn, and Eytan Ruppin. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nature Medicine*, 24(10):1545–1549, 2018.

[15] M.H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, and A. Weerasinghe. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 174:1034–1035.

[16] O.B. Bakker, C.J. Xu, H.J.P.M. Koenen, L.A.B. Joosten, and M.G. Netea. Deconvolution of bulk blood eqtl effects into immune cell subpopulations. *BMC Bioinformatics*, 21(1).

[17] Albert László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[18] Amorette Barber. Costimulation of Effector CD8+ T Cells: Which Receptor is Optimal for Immunotherapy? *MOJ Immunology*, 1(2), 2014.

[19] Ruth Barshir, Omer Basha, Amir Eluk, Ilan Y. Smoly, Alexander Lan, and Esti Yeger-Lotem. The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Research*, 41(D1), 2013.

[20] Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, Marc E. Martignoni, Angelika Werner, Rüdiger Hein, Dirk H. Busch, Christian Peschel, Roland Rad, Jürgen Cox, Matthias Mann, and Angela M. Krackhardt. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature Communications*, 7, 2016.

[21] U. Ben-David and A. Amon. Context is everything: aneuploidy in cancer. *Nat. Rev. Genet*, 21:44–62.

[22] R. Beroukhim, C.H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, and J. Donovan. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463:899–905.

[23] D.D. Billadeau and P.J. Leibson. Itams versus itims: striking a balance during cell regulation. *J Clin Invest*, 109(2):161–168.

[24] A. Bird. DNA methylation patterns and epigenetic memory, 2002.

[25] Dima Blokh, Danny Segev, and Roded Sharan. The approximability of shortest path-based graph orientations of protein–protein interaction networks. *Journal of Computational Biology*, 20(12):945–957, 2013.

[26] Russell Bonneville, Melanie A. Krook, Esko A. Kautto, Jharna Miya, Michele R. Wing, Hui-Zi Chen, Julie W. Reeser, Lianbo Yu, and Sameek Roychowdhury. Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precision Oncology*, (1):1–15, 2017.

[27] J. Borst, T. Ahrends, N. Babała, C.J.M. Melief, and W. Kastenmüller. Cd4+ t cell help in cancer immunology and immunotherapy. *Nature Reviews Immunology*, 18:635–647.

[28] Michael M. Boyiadzis, John M. Kirkwood, John L. Marshall, Colin C. Pritchard, Nilofer S. Azad, and James L. Gulley. Significance and implications of FDA approval of pembrolizumab for biomarker-defined disease, 2018.

[29] R. Braun, S. Ronquist, D. Wangsa, H. Chen, L. Anthuber, and T. Gemoll. Single chromosome aneuploidy induces genome-wide perturbation of nuclear organization and gene expression. *Neoplasia*, 21:401–412.

[30] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[31] Ashton Breitkreutz, Hyungwon Choi, Jeffrey R Sharom, Lorrie Boucher, Victor Neduva, Brett Larsen, Zhen-Yuan Lin, Bobby-Joe Breitkreutz, Chris Stark, Guomin Liu, et al. A global protein kinase and phosphatase interaction network in yeast. *Science*, 328(5981):1043–1046, 2010.

[32] R. Bro and S. De Jong. A fast non-negativity-constrained least squares algorithm.

[33] J.M. Brown, L. Recht, and S. Strober. The promise of targeting macrophages in cancer therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 23:3241–3250.

[34] T. Budden, R.J. Davey, R.E. Vilain, K.A. Ashton, S.G. Braye, N.J. Beveridge, and N.A. Bowden. Repair of uvb-induced dna damage is reduced in melanoma due to low xpc and global genome repair. *Oncotarget*, 7:60940–60953.

[35] K.S. Campbell and A.K. Purdy. Structure/function of human killer cell immunoglobulin-like receptors: Lessons from polymorphisms, evolution, crystal structures and mutations. *Immunology*, 132(3):315–325.

[36] J. Camps, J.J. Pitt, G. Emons, A.B. Hummon, C.M. Case, and M. Grade. Genetic amplification of the notch modulator lnx2 upregulates the wnt/beta-catenin pathway in colorectal cancer. *Cancer Res*, 73:2003–2013.

[37] Jude Canon, Rebecca Bryant, Martine Roudier, Daniel G. Branstetter, and William C. Dougall. RANKL inhibition combined with tamoxifen treatment increases anti-tumor efficacy and prevents tumor-induced bone destruction in an estrogen receptor-positive breast cancer bone metastasis model. *Breast Cancer Research and Treatment*, 135(3):771–780, 2012.

[38] S.L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, and T. Zack. Absolute quantification of somatic dna alterations in human cancer. *Nat. Biotechnol*, 30:413–421.

[39] Ankur Chakravarthy, Andrew Furness, Kroopa Joshi, Ehsan Ghorani, Kirsty Ford, Matthew J. Ward, Emma V. King, Matt Lechner, Teresa Marafioti, Sergio A. Quezada, Gareth J. Thomas, Andrew Feber, and Tim R. Fenton. Pan-cancer deconvolution of tumour composition using DNA methylation. *Nature Communications*, 9(1), 2018.

[40] H. Chen, Q.Y. Weng, and D.E. Fisher. Uv signaling pathways within the skin. *The Journal of investigative dermatology*, 134:2080–2085.

[41] L. Chen and D.B. Flies. Molecular mechanisms of t cell co-stimulation and co-inhibition. *Nat Rev Immunol*, 13(4):227–242.

[42] Shihao Chen, Li Fen Lee, Timothy S. Fisher, Bart Jessen, Mark Elliott, Winston Evering, Kathryn Logronio, Guang Huan Tu, Konstantinos Tsaparikos, Xiaoai Li, Hui Wang, Chi Ying, Mengli Xiong, Todd Van Arsdale, and John C. Lin. Combination of 4-1BB agonist and PD-1 antagonist promotes antitumor effector/memory CD8 T cells in a Poorly Immunogenic Tumor Model. *Cancer Immunology Research*, 3(2):149–160, 2015.

[43] Cariad Chester, Miguel F. Sanmamed, Jun Wang, and Ignacio Melero. Immunotherapy targeting 4-1BB: Mechanistic rationale, clinical results, and future strategies, 2018.

[44] Jen Tsan Chi, Edwin H. Rodriguez, Zhen Wang, Dimitry S.A. Nuyten, Sayan Mukherjee, Matt Van De Rijn, Marc J. Van De Vijver, Trevor Hastie, and Patrick O. Brown. Gene expression programs of human smooth muscle cells: Tissue-specific differentiation and prognostic significance in breast cancers. *PLoS Genetics*, 3(9):1770–1784, 2007.

[45] N. Chihara, A. Madi, T. Kondo, H. Zhang, N. Acharya, M. Singer, J. Nyman, N.D. Marjanovic, M.S. Kowalczyk, and C. Wang. Induction and transcriptional regulation of the co-inhibitory gene module in t cells. *Nature*, 558:454–459.

[46] Yeonjoo Choi, Yaoyao Shi, Cara L. Haymaker, Aung Naing, Gennaro Ciliberto, and Joud Hajjar. T-cell agonists in cancer immunotherapy, 2020.

[47] Garry Choy, Peter Choyke, and Steven K Libutti. Current advances in molecular imaging: noninvasive in vivo bioluminescent and fluorescent optical imaging in cancer research. *Molecular imaging*, 2(4):15353500200303142, 2003.

[48] Allison S. Cleary, Travis L. Leonard, Shelley A. Gestl, and Edward J. Gunther. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature*, 508(1):113–117, 2014.

[49] Seth B. Coffelt, Kelly Kersten, Chris W. Doornebal, Jorieke Weiden, Kim Vrijland, Cheei Sing Hau, Niels J.M. Verstegen, Metamia Ciampricotti, Lukas J.A.C. Hawinkels, Jos Jonkers, and Karin E. De Visser. IL-17-producing γ δ T cells and neutrophils conspire to promote breast cancer metastasis. *Nature*, 522(7556):345–348, 2015.

[50] C.J. Cohen, J.J. Gartner, M. Horovitz-Fried, K. Shamalov, K. Trebska-McGowan, V.V. Bliskovsky, M.R. Parkhurst, C. Ankri, T.D. Prickett, and J.S. Crystal. Isolation of neoantigen-specific t cells from tumor and peripheral lymphocytes. *The Journal of clinical investigation*, 125:3981–3991.

[51] Ezra E.W. Cohen, Michael J. Pishvaian, Dale R. Shepard, Ding Wang, Jared Weiss, Melissa L. Johnson, Christine H. Chung, Ying Chen, Bo Huang, Craig B. Davis, Francesca Toffalorio, Aron Thall, and Steven F. Powell. A phase Ib study of utomilumab (PF-05082566) in combination with mogamulizumab in patients with advanced solid tumors. *Journal for ImmunoTherapy of Cancer*, 7(1), 2019.

[52] Isidro Cortes-Ciriano, Sejoon Lee, Woong Yang Park, Tae Min Kim, and Peter J. Park. A molecular portrait of microsatellite instability across multiple cancers. *Nature Communications*, 8, 2017.

[53] Christina Curtis et. al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[54] H. X. Dang, B. S. White, S. M. Foltz, C. A. Miller, J. Luo, R. C. Fields, and C. A. Maher. ClonEvol: Clonal ordering and visualization in cancer sequencing. *Annals of Oncology*, 28(12):3076–3082, 2017.

[55] T. Davoli, A.W. Xu, K.E. Mengwasser, L.M. Sack, J.C. Yoon, P.J. Park, and S.J. Elledge. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155:948–962.

[56] Teresa Davoli, Hajime Uno, Eric C. Wooten, and Stephen J. Elledge. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, 355(6322), 2017.

[57] Edward C. De Fabo, Frances P. Noonan, Thomas Fears, and Glenn Merlino. Ultraviolet B but not ultraviolet A radiation initiates melanoma. *Cancer Research*, 64(18):6372–6376, 2004.

[58] Álvaro de Mingo Pulido, Alycia Gardner, Shandi Hiebler, Hatem Soliman, Hope S. Rugo, Matthew F. Krummel, Lisa M. Coussens, and Brian Ruffell. TIM-3 Regulates CD103+ Dendritic Cell Function and Response to Chemotherapy in Breast Cancer. *Cancer Cell*, 33(1):60–74.e6, 2018.

[59] Peter C. DeWeirdt, Kendall R. Sanson, Annabel K. Sangree, Mudra Hegde, Ruth E. Hanna, Marissa N. Feeley, Audrey L. Griffith, Teng Teng, Samantha M. Borys, Christine Strand, J. Keith Joung, Benjamin P. Kleinstiver, Xuewen Pan, Alan Huang, and John G. Doench. Optimization of AsCas12a for combinatorial genetic screens in human cells. *Nature Biotechnology*, 39(1):94–104, 2021.

[60] M. Dmitrijeva, S. Ossowski, L. Serrano, and M.H. Schaefer. Tissue-specific dna methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. *Nucleic Acids Res*, 46:7022–7039.

[61] M. Dong, A. Thennavan, and E. Urrutia. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Brief Bioinform*, 22(1):416–427.

[62] M. Durrbaum and Z. Storchova. Effects of aneuploidy on gene expression: implications for cancer. *FEBS J*, 283:791–802.

[63] W. Erskine, I. Kusmenoglu, F. J. Muehlbauer, and R. J. Summerfield. Breeding for increased biomass and persistent crop residues in cool-season food legumes. *Genetics*, 148(4):191–197, 2000.

[64] Celine Everaert, Manuel Luypaert, Jesper L.V. Maag, Quek Xiu Cheng, Marcel E. DInger, Jan Hellemans, and Pieter Mestdagh. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Scientific Reports*, 7(1), 2017.

[65] Natalie S. Fox, Syed Haider, Adrian L. Harris, and Paul C. Boutros. Landscape of transcriptomic interactions between breast cancer and its microenvironment. *Nature Communications*, 10(1), 2019.

[66] et. al Frederick Acre Vargas. Fc-Optimized Anti-CD25 Depletes Tumor-Infiltrating Regulatory T Cells and Synergizes with PD-1 Blockade to Eradicate Established Tumors. *Immunity*, 46(4):577–586, 2017.

[67] William K. Funkhouser, Ira M. Lubin, Federico A. Monzon, Barbara A. Zehnbauer, James P. Evans, Shuji Ogino, and Jan A. Nowak. Relevance, Pathogenesis, and testing algorithm for mismatch repair-defective colorectal carcino-

mas: A report of the association for molecular pathology. *Journal of Molecular Diagnostics*, 14(2):91–103, 2012.

[68] Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, Franck Zinzindohoué, Patrick Bruneval, Paul Henri Cugnenc, Zlatko Trajanoski, Wolf Herman Fridman, and Franck Pagès. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964, 2006.

[69] R. Gaujoux and C. Seoighe. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. *Infect Genet Evol*, 12(5):913–921.

[70] J. M. Gee, J. F. Robertson, E. Gutteridge, I. O. Elis, S. E. Pinder, M. Rubini, and R. I. Nicholson. Epidermal growth factor receptor/HER2/insulin-like growth factor receptor signalling and oestrogen receptor activity in clinical breast cancer. *Endocrine-Related Cancer*, 12(SUPPL. 1):S99–S111, 2005.

[71] R.S. Gejman, A.Y. Chang, H.F. Jones, K. DiKun, A.A. Hakimi, A. Schietinger, and D.A. Scheinberg. Rejection of immunogenic tumor clones is limited by clonal fraction. eLife 7.

[72] Andrew J. Gentles, Angela Bik Yu Hui, Weiguo Feng, Armon Azizi, Ramesh V. Nair, Gina Bouchard, David A. Knowles, Alice Yu, Youngtae Jeong, Alborz Bejnood, Erna Forgó, Sushama Varma, Yue Xu, Amanda Kuong, Viswam S. Nair, Rob West, Matt Van De Rijn, Chuong D. Hoang, Maximilian Diehn, and Sylvia K. Plevritis. A human lung tumor microenvironment interactome identifies clinically relevant cell-type cross-talk. *Genome Biology*, 21(1), 2020.

[73] G. Germano, S. Lamba, G. Rospo, L. Barault, A. Magri, F. Maione, M. Russo, G. Crisafulli, A. Bartolini, and G. Lerda. Inactivation of dna repair triggers neoantigen generation and impairs tumour growth. *Nature*, 552:116–120.

[74] Umesh Ghoshdastider, Marjan Mojtabavi Naeini, Neha Rohatgi, Egor Revkov, Angeline Wong, Sundar Solai, Tin Trung Nguyen, Joe Yeong, Jabed Iqbal, Puay Hoon Tan, Balram Chowbay, Ramanuj DasGupta, and Anders Jacobsen Skanderup. Data-driven inference of crosstalk in the tumor microenvironment. *bioRxiv*, page 835512, jan 2019.

[75] Tuba N. Gide, Camelia Quek, Alexander M. Menzies, Annie T. Tasker, Ping Shang, Jeff Holst, Jason Madore, Su Yin Lim, Rebecca Velickovic, Matthew Wongchenko, Yibing Yan, Serigne Lo, Matteo S. Carlino, Alexander Guminski, Robyn P.M. Saw, Angel Pang, Helen M. McGuire, Umaimainthan Palendira, John F. Thompson, Helen Rizos, Ines Pires da Silva, Marcel Batten, Richard A. Scolyer, Georgina V. Long, and James S. Wilmott. Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and

Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell*, 35(2):238–255.e6, 2019.

[76] Jesse Gillis and Paul Pavlidis. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3), 2012.

[77] Mary J. Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N. Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the Xena platform, 2020.

[78] Thomas Gonatopoulos-Pournatzis, Michael Aregger, Kevin R. Brown, Shaghayegh Farhangmehr, Ulrich Braunschweig, Henry N. Ward, Kevin C.H. Ha, Alexander Weiss, Maximilian Billmann, Tanja Durbic, Chad L. Myers, Benjamin J. Blencowe, and Jason Moffat. Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9–Cas12a platform. *Nature Biotechnology*, 38(5):638–648, 2020.

[79] T. Gong and J.D. Szustakowski. Deconrnaseq: A statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085.

[80] C.M. Gonçalves, S.N. Henriques, R.F. Santos, and A.M. Carmo. Cd6, a rheostat-type signalosome that tunes t cell activation. *Front Immunol*, 9.

[81] N.A. Graham, A. Minasyan, A. Lomova, A. Cass, N.G. Balanis, and M. Friedman. Recurrent patterns of dna copy number alterations in tumors reflect metabolic selection pressures. *Mol. Syst. Biol*, 13(914).

[82] E. Gronroos and C. Lopez-Garcia. Tolerance of chromosomal instability in cancer: Mechanisms and therapeutic opportunities. *Cancer Res*, 78:6529–6535.

[83] A. Gros, M.R. Parkhurst, E. Tran, A. Pasetto, P.F. Robbins, S. Ilyas, T.D. Prickett, J.J. Gartner, J.S. Crystal, and I.M. Roberts. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nature medicine*, 22:433–438.

[84] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors. Circlize implements and enhances circular visualization in r. *Bioinformatics*, 30(19):2811–2812.

[85] M.M. Gubin, M.N. Artyomov, E.R. Mardis, and R.D. Schreiber. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *The Journal of clinical investigation*, 125:3413–3421.

[86] M.M. Gubin, E. Esaulova, J.P. Ward, O.N. Malkova, D. Runci, P. Wong, T. Noguchi, C.D. Arthur, W. Meng, and E. Alspach. High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. *Cell*, 175:1014–1030 1019.

[87] M.M. Gubin, X. Zhang, H. Schuster, E. Caron, J.P. Ward, T. Noguchi, Y. Ivanova, J. Hundal, C.D. Arthur, and W.J. Krebber. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*, 515:577–581.

[88] C. Guillerey, N.D. Huntington, and M.J. Smyth. Targeting natural killer cells in cancer immunotherapy. *Nature Immunology*, 17:1025.

[89] Carino Gurjao, Dina Tsukrov, Maxim Imakaev, Lovelace J. Luquette, and Leonid A. Mirny. Limited evidence of tumour mutational burden as a biomarker of response to immunotherapy. *bioRxiv*, page 2020.09.03.260265, jan 2020.

[90] O.A.W. Haabeth, A.A. Tveita, M. Fauskanger, F. Schjesvold, K.B. Lorvik, P.O. Hofgaard, H. Omholt, L.A. Munthe, Z. Dembic, and A. Corthay. How do cd4+ t cells detect and eliminate tumor cells that either lack or express mhc class ii molecules? *Frontiers in Immunology*, 5.

[91] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje Van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774, 2012.

[92] Tamir Hazan and Tommi Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 991–998, 2012.

[93] S. Heim and F. Mitelman. *Cancer Cytogenetics*. John Wiley Sons, Hoboken.

[94] Ellen Heitzer, Imran S Haque, Charles ES Roberts, and Michael R Speicher. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics*, 20(2):71–88, 2019.

[95] M.D. Hellmann, T. Nathanson, H. Rizvi, B.C. Creelan, F. Sanchez-Vega, A. Ahuja, A. Ni, J.B. Novik, L.M.B. Mangarin, and M. Abu-Akeel. *Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer*. Cancer Cell.

[96] Luisa Henkel, Benedikt Rauscher, and Michael Boutros. Context-dependent genetic interactions in cancer, 2019.

[97] Shay Houri and Roded Sharan. Sign assignment problems on protein networks. In *WABI*, pages 338–345. Springer, 2012.

[98] Xiu Huang, David F Stern, and Hongyu Zhao. Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival–evidence from tcga pan-cancer data. *Scientific reports*, 6(1):1–9, 2016.

[99] W. Hugo, J.M. Zaretsky, L. Sun, C. Song, B.H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, and G. Cherry. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165:35–44.

[100] Las Rivas J and Fontanillo C. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):1–8.

[101] J. J. James, A. J. Evans, S. E. Pinder, E. Gutteridge, K. L. Cheung, S. Chan, and J. F.R. Robertson. Bone metastases from breast carcinoma: Histopathological-radiological correlations and prognostic features. *British Journal of Cancer*, 89(4):660–665, 2003.

[102] Michalina Janiszewska, Doris P. Tabassum, Zafira Castaño, Simona Cristea, Kimiyo N. Yamamoto, Natalie L. Kingston, Katherine C. Murphy, Shaokun Shu, Nicholas W. Harper, Carlos Gil Del Alcazar, Maša Alečković, Muhammad B. Ekram, Ofir Cohen, Minsuk Kwak, Yuanbo Qin, Tyler Laszewski, Adrienne Luoma, Andriy Marusyk, Kai W. Wucherpfennig, Nikhil Wagle, Rong Fan, Franziska Michor, Sandra S. McAllister, and Kornelia Polyak. Subclonal cooperation drives metastasis by modulating local and systemic immune microenvironments. *Nature Cell Biology*, 21(7):879–888, 2019.

[103] Livnat Jerby-Arnon, Parin Shah, Michael S. Cuoco, Christopher Rodman, Mei Ju Su, Johannes C. Melms, Rachel Leeson, Abhay Kanodia, Shaolin Mei, Jia Ren Lin, Shu Wang, Bokang Rabasha, David Liu, Gao Zhang, Claire Margolais, Orr Ashenberg, Patrick A. Ott, Elizabeth I. Buchbinder, Rizwan Haq, F. Stephen Hodi, Genevieve M. Boland, Ryan J. Sullivan, Dennie T. Frederick, Benchun Miao, Tabea Moll, Keith T. Flaherty, Meenhard Herlyn, Russell W. Jenkins, Rohit Thummalapalli, Monika S. Kowalczyk, Israel Cañadas, Bastian Schilling, Adam N.R. Cartwright, Adrienne M. Luoma, Shruti Malu, Patrick Hwu, Chantale Bernatchez, Marie Andrée Forget, David A. Barbie, Alex K. Shalek, Itay Tirosh, Peter K. Sorger, Kai Wucherpfennig, Eliezer M. Van Allen, Dirk Schadendorf, Bruce E. Johnson, Asaf Rotem, Orit Rozenblatt-Rosen, Levi A. Garraway, Charles H. Yoon, Benjamin Izar, and Aviv Regev. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell*, 175(4):984–997.e24, 2018.

[104] B. Jew, M. Alvarez, and E. Rahmani. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun*, 11(1).

[105] Peng Jiang, Shengqing Gu, Deng Pan, Jingxin Fu, Avinash Sahu, Xihao Hu, Ziyi Li, Nicole Traugh, Xia Bu, Bo Li, Jun Liu, Gordon J. Freeman, Myles A. Brown, Kai W. Wucherpfennig, and X. Shirley Liu. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nature Medicine*, 24(10):1550–1558, 2018.

[106] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey, 2019.

[107] D. Holstead Jones, Tomoki Nakashima, Otto H. Sanchez, Ivona Kozieradzki, Svetlana V. Komarova, Ildiko Sarosi, Sean Morony, Evelyn Rubin, Renu Sarao, Carlo V. Hojilla, Vukoslav Komnenovic, Young Yun Kong, Martin Schreiber, S. Jeffrey Dixon, Stephen M. Sims, Rama Khokha, Teiji Wada, and Josef M. Penninger. Regulation of cancer cell migration and bone metastasis by RANKL. *Nature*, 440(7084):692–696, 2006.

[108] Shelly Kalaora, Yochai Wolf, Tali Feferman, Eilon Barnea, Erez Greenstein, Dan Reshef, Itay Tirosh, Alexandre Reuben, Sushant Patkar, Ronen Levy, Juliane Quinkhardt, Tana Omokoko, Nouar Qutob, Ofra Golani, Jianhua Zhang, Xizeng Mao, Xingzhi Song, Chantale Bernatchez, Cara Haymaker, Marie Andrée Forget, Caitlin Creasy, Polina Greenberg, Brett W. Carter, Zachary A. Cooper, Steven A. Rosenberg, Michal Lotem, Ugur Sahin, Guy Shakhar, Eytan Ruppin, Jennifer A. Wargo, Nir Friedman, Arie Admon, and Yardena Samuels. Combined analysis of antigen presentation and T-cell recognition reveals restricted immune responses in melanoma. *Cancer Discovery*, 8(11):1366–1375, 2018.

[109] Kumaran Kandasamy, S. Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S. Sameer Kumar, Abhilash K. Venugopal, Deepthi Telikicherla, Daniel J. Navarro, Suresh Mathivanan, Christian Pecquet, Sashi Kanth Gollapudi, Sudhir Gopal Tattikota, Shyam Mohan, Hariprasad Padhukasahasram, Yashwanth Subbannayya, Renu Goel, Harrys K.C. Jacob, Jun Zhong, Raja Sekhar, Vishalakshi Nanjappa, Lavanya Balakrishnan, Roopashree Subbaiah, Y. L. Ramachandra, B. Abdul Rahiman, T. S. Keshava Prasad, Jian Xin Lin, Jon C.D. Houtman, Stephen Desiderio, Jean Christophe Renauld, Stefan Constantinescu, Osamu Ohara, Toshio Hirano, Masato Kubo, Sujay Singh, Purvesh Khatri, Sorin Draghici, Gary D. Bader, Chris Sander, Warren J. Leonard, and Akhilesh Pandey. NetPath: A public resource of curated signal transduction pathways. *Genome Biology*, 11(1):R3, 2010.

[110] M. Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Sci*, 28(11):1947–1951.

[111] M. Kanehisa and Goto S. KEGG. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

[112] J.N. Kather, M. Suarez-Carmona, P. Charoentong, C.A. Weis, D. Hirsch, P. Bankhead, M. Horning, D. Ferber, I. Kel, and E. Herpel. Topography of cancer-associated immune cells in human solid tumors. eLife 7.

[113] Patrick Kemmeren, Katrin Sameith, Loes AL van de Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

[114] Hoon Kim, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D Gujar, Jens Luebeck, Jihe Liu, Viraj Deshpande, Utkrisht Rajkumar, Sandeep Namburi, et al. Extrachromosomal dna is associated with oncogene amplification and poor outcome across multiple cancers. *Nature Genetics*, 52(9):891–897, 2020.

[115] T. Knutsen, V. Gobu, R. Knaus, H. Padilla-Nash, M. Augustus, R.L. Strausberg, I.R. Kirsch, K. Sirotkin, and T. Ried. The interactive online sky/m-fish cgh database and the entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes Cancer*, 44:52–64.

[116] S. Knuutila, K. Autio, and Y. Aalto. Online access to cgh data of dna sequence copy number changes. *Am. J. Pathol*, 157(689).

[117] L. Konig, F.D. Mairinger, O. Hoffmann, A.K. Bittner, K.W. Schmid, R. Kimmig, S. Kasimir-Bauer, and A. Bankfalvi. Dissimilar patterns of tumor-infiltrating immune cells at the invasive tumor front and tumor center are associated with response to neoadjuvant chemotherapy in primary breast cancer. *BMC cancer*, 19:120.

[118] S. Kurtulus, A. Madi, G. Escobar, M. Klapholz, J. Nyman, E. Christian, M. Pawlak, D. Dionne, J. Xia, and O. Rozenblatt-Rosen. Checkpoint blockade immunotherapy induces dynamic changes in pd-1(-)cd8(+) tumor-infiltrating t cells. *Immunity*, 50:181–194 186.

[119] John E Ladbury and Stefan T Arold. Noise in cellular signaling pathways: causes and effects. *Trends in biochemical sciences*, 37(5):173–178, 2012.

[120] S.A. Lambert, A. Jolma, and L.F. Campitelli. The human transcription factors. *Cell*, 172(4):650–665.

[121] J. Larkin, V. Chiarion-Sileni, R. Gonzalez, J.J. Grob, C.L. Cowey, C.D. Lao, D. Schadendorf, R. Dummer, M. Smylie, and P. Rutkowski. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *The New England journal of medicine*, 373:23–34.

[122] Dung T. Le, Jennifer N. Uram, Hao Wang, Bjarne R. Bartlett, Holly Kemberling, Aleksandra D. Eyring, Andrew D. Skora, Brandon S. Luber, Nilofer S. Azad, Dan Laheru, Barbara Biedrzycki, Ross C. Donehower, Atif Zaheer, George A. Fisher, Todd S. Crocenzi, James J. Lee, Steven M. Duffy, Richard M. Goldberg, Albert de la Chapelle, Minori Koshiji, Feriyl Bhaijee, Thomas Huebner, Ralph H. Hruban, Laura D. Wood, Nathan Cuka, Drew M. Pardoll, Nickolas Papadopoulos, Kenneth W. Kinzler, Shibin Zhou, Toby C. Cornish, Janis M. Taube, Robert A. Anders, James R. Eshleman, Bert Vogelstein, and Luis A. Diaz. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, 372(26):2509–2520, 2015.

[123] Insuk Lee, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121, 2011.

[124] T.I. Lee and R.A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251.

[125] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, and J.D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883.

[126] Bo Li and Jun Z. Li. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome biology*, 15(9):473, 2014.

[127] H. Li, A.M. Leun, I. Yofe, Y. Lubling, D. Gelbard-Solodkin, A.C.J. Akkooi, M. Braber, E.A. Rozeman, J. Haanen, and C.U. Blank. Dysfunctional cd8 t cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell*, 176:775–789 718.

[128] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[129] Taiwen Li, Jingyu Fan, Binbin Wang, Nicole Traugh, Qianming Chen, Jun S. Liu, Bo Li, and X. Shirley Liu. TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Research*, 77(21):e108–e110, 2017.

[130] Carsten Linnemann, Marit M. Van Buuren, Laura Bies, Els M.E. Verdegaal, Remko Schotte, Jorg J.A. Calis, Sam Behjati, Arno Velds, Henk Hilkmann, Dris El Atmioui, Marten Visser, Michael R. Stratton, John B.A.G. Haanen, Hergen Spits, Sjoerd H. Van Der Burg, and Ton N.M. Schumacher. High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nature Medicine*, 21(1):81–85, 2015.

[131] David Liu, Bastian Schilling, Derek Liu, Antje Sucker, Elisabeth Livingstone, Livnat Jerby-Amon, Lisa Zimmer, Ralf Gutzmer, Imke Satzger, Carmen Loquai, Stephan Grabbe, Natalie Vokes, Claire A. Margolis, Jake Conway, Meng Xiao He, Haitham Elmarakeby, Felix Dietlein, Diana Miao, Adam Tracy, Helen Gogas, Simone M. Goldinger, Jochen Utikal, Christian U. Blank, Ricarda Rauschenberg, Dagmar von Bubnoff, Angela Krackhardt, Benjamin Weide, Sebastian Haferkamp, Felix Kiecker, Ben Izar, Levi Garraway, Aviv Regev, Keith Flaherty, Annette Paschen, Eliezer M. Van Allen, and Dirk Schadendorf. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nature Medicine*, 25(12):1916–1927, 2019.

[132] Nicolas J. Llosa, Michael Cruise, Ada Tam, Elizabeth C. Wicks, Elizabeth M. Hechenbleikner, Janis M. Taube, Richard L. Blosser, Hongni Fan, Hao Wang, Brandon S. Luber, Ming Zhang, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, Cynthia L. Sears, Robert A. Anders, Drew M. Pardoll, and Franck Housseau. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discovery*, 5(1):43–51, 2015.

[133] Lawrence A. Loeb. Mutator Phenotype May Be Required for Multistage Carcinogenesis. *Cancer Research*, 51(12):3075–3079, 1991.

[134] Ulrike Löhr and Leslie Pick. Cofactor-interaction motifs and the cooption of a homeotic Hox protein into the segmentation pathway of Drosophila melanogaster. *Current Biology*, 15(7):643–649, 2005.

[135] Lichun Ma, Maria O. Hernandez, Yongmei Zhao, Monika Mehta, Bao Tran, Michael Kelly, Zachary Rae, Jonathan M. Hernandez, Jeremy L. Davis, Sean P. Martin, David E. Kleiner, Stephen M. Hewitt, Kris Ylaya, Bradford J. Wood, Tim F. Greten, and Xin Wei Wang. Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. *Cancer Cell*, 36(4):418–430.e6, 2019.

[136] Kenzie D MacIsaac, Ting Wang, D Benjamin Gordon, David K Gifford, Gary D Stormo, and Ernest Fraenkel. An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC bioinformatics*, 7(1):113, 2006.

[137] Helder Maiato and Elsa Logarinho. Mitotic spindle multipolarity without centrosome amplification. *Nature Cell Biology*, 16(5):386–394, 2014.

[138] Edward M. Marcotte, Matteo Pellegrini, Michael J. Thompson, Todd O. Yeates, and David Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, nov 1999.

[139] Andriy Marusyk, Michalina Janiszewska, and Kornelia Polyak. Intratumor heterogeneity: The rosetta stone of therapy resistance. *Cancer cell*, 37(4):471–484, 2020.

[140] Andriy Marusyk, Doris P. Tabassum, Philipp M. Altrock, Vanessa Almendro, Franziska Michor, and Kornelia Polyak. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*, 514(7520):54–58, 2014.

[141] Tarik F Massoud and Sanjiv S Gambhir. Integrating noninvasive molecular imaging into molecular medicine: an evolving paradigm. *Trends in molecular medicine*, 13(5):183–191, 2007.

[142] K.A. McDonald, T. Kawaguchi, Q. Qi, X. Peng, M. Asaoka, J. Young, M. Opyrchal, L. Yan, S. Patnaik, and E. Otsuji. Tumor heterogeneity correlates with less immune response and worse survival in breast cancer patients. annals of surgical oncology.

[143] N. McGranahan, A.J. Furness, R. Rosenthal, S. Ramskov, R. Lyngaa, S.K. Saini, M. Jamal-Hanjani, G.A. Wilson, N.J. Birkbak, and C.T. Hiley. Clonal neoantigens elicit t cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351:1463–1469.

[144] N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168:613–628.

[145] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[146] Mitchell Melanie. An introduction to genetic algorithms By Melanie Mitchell. MIT Press, Cambridge, MA. (1996). 205 pages. $30.00. *Computers & Mathematics with Applications*, 32(6):133, 1996.

[147] Ignacio Melero, Daniel Hirschhorn-Cymerman, Aizea Morales-Kastresana, Miguel F. Sanmamed, and Jedd D. Wolchok. Agonist antibodies to TNFR molecules that costimulate T and NK cells. *Clinical Cancer Research*, 19(5):1044–1053, 2013.

[148] D. Miao, C.A. Margolis, W. Gao, M.H. Voss, W. Li, D.J. Martini, C. Norton, D. Bosse, S.M. Wankowicz, and D. Cullen. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science*, 359:801–806.

[149] D. Miao, C.A. Margolis, N.I. Vokes, D. Liu, A. Taylor-Weiner, S.M. Wankowicz, D. Adeegbe, D. Keliher, B. Schilling, and A. Tracy. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nature genetics*, 50:1271–1281.

[150] D. Miao, C.A. Margolis, N.I. Vokes, D. Liu, A. Taylor-Weiner, S.M. Wankowicz, D. Adeegbe, D. Keliher, B. Schilling, and A. Tracy. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nature genetics*, 50:1271–1281.

[151] Christopher A. Miller, Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, Michael H. Tomasson, Timothy A. Graubert, Matthew J. Walter, Matthew J. Ellis, William Schierding, John F. DiPersio, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology*, 10(8), 2014.

[152] Idan Milo, Marie Bedora-Faure, Zacarias Garcia, Ronan Thibaut, Leïla Périé, Guy Shakhar, Ludovic Deriano, and Philippe Bousso. The immune system profoundly restricts intratumor genetic heterogeneity. *Science Immunology*, 3(29), 2018.

[153] Marja Moerkens, Yinghui Zhang, Lynn Wester, Bob van de Water, and John H.N. Meerman. Epidermal growth factor receptor signalling in human breast cancer cells operates parallel to estrogen receptor α signalling and results in tamoxifen insensitive proliferation. *BMC Cancer*, 14(1):283, 2014.

[154] S. Mohammadi, N. Zuckerman, A. Goldsmith, and A. Grama. A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc IEEE*, 105(2):340–366.

[155] G. Monaco, B. Lee, and W. Xu. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell Rep*, 26(6):1627–1640.

[156] L.G. Morris, N. Riaz, A. Desrichard, Y. Senbabaoglu, A.A. Hakimi, V. Makarov, J.S. Reis-Filho, and T.A. Chan. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 7:10051–10063.

[157] Melody K Morris, Julio Saez-Rodriguez, Peter K Sorger, and Douglas A Lauffenburger. Logic-based models for the analysis of cell signaling networks. *Biochemistry*, 49(15):3216–3224, 2010.

[158] Pawel Muranski, Andrea Boni, Paul A. Antony, Lydie Cassard, Kari R. Irvine, Andrew Kaiser, Chrystal M. Paulos, Douglas C. Palmer, Christopher E. Touloukian, Krzysztof Ptak, Luca Gattinoni, Claudia Wrzesinski, Christian S. Hinrichs, Keith W. Kerstann, Lionel Feigenbaum, Chi Chao Chan, and Nicholas P. Restifo. Tumor-specific Th17-polarized cells eradicate large established melanoma. *Blood*, 112(2):362–373, 2008.

[159] K. Murphy and C. Weaver. Janeway's immunbiology.

[160] Donna M. Muzny et. al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[161] S. Myllykangas, T. Bohling, and S. Knuutila. Specificity, selection and significance of gene amplifications in cancer. semin. *Cancer Biol*, 17:42–55.

[162] Aaron M. Newman, Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, 2015.

[163] Aaron M. Newman, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn, and Ash A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019.

[164] A.M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, C.D. Hoang, M. Diehn, and A.A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12:453.

[165] Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

[166] O.E. Ogundijo and X. Wang. A sequential monte carlo approach to gene expression deconvolution. *PLoS One*, 12(10).

[167] A.B. Olshen, E.S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5:557–572.

[168] V. Onuchic, R.J. Hartmaier, and D.N. Boone. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell Rep*, 17(8):2075–2086.

[169] Stephen Paget. the Distribution of Secondary Growths in Cancer of the Breast. *The Lancet*, 133(3421):571–573, 1889.

[170] S.J. Patel, N.E. Sanjana, R.J. Kishton, A. Eidizadeh, S.K. Vodnala, M. Cam, J.J. Gartner, L. Jia, S.M. Steinberg, and T.N. Yamamoto. Identification of essential genes for cancer immunotherapy. *Nature*, 548:537–542.

[171] Sushant Patkar, Assaf Magen, Roded Sharan, and Sridhar Hannenhalli. A network diffusion approach to inferring sample-specific function reveals functional changes associated with breast cancer. *PLoS computational biology*, 13(11):e1005793, 2017.

[172] Sushant Patkar and Roded Sharan. An optimization framework for network annotation. *Bioinformatics*, 34(13):i502–i508, 2018.

[173] Michael J. Penciana and Ralph B. D'Agostino. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004.

[174] D. Pende, M. Falco, and M. Vitale. Killer ig-like receptors (kirs): Their role in nk cell modulation and developments leading to their clinical exploitation. *Front Immunol*, 10(MAY).

[175] Xianlu Laura Peng, Richard A. Moffitt, Robert J. Torphy, Keith E. Volmar, and Jen Jen Yeh. De novo compartment deconvolution and weight estimation of tumor samples using DECODER. *Nature Communications*, 10(1), 2019.

[176] Eva Pérez-Guijarro, Howard H. Yang, Romina E. Araya, Rajaa El Meskini, Helen T. Michael, Suman Kumar Vodnala, Kerrie L. Marie, Cari Smith, Sung Chin, Khiem C. Lam, Andres Thorkelsson, Anthony J. Iacovelli, Alan Kulaga, Anyen Fon, Aleksandra M. Michalowski, Willy Hugo, Roger S. Lo, Nicholas P. Restifo, Shyam K. Sharan, Terry Van Dyke, Romina S. Goldszmid, Zoe Weaver Ohler, Maxwell P. Lee, Chi Ping Day, and Glenn Merlino. Multimodel preclinical platform predicts clinical response of melanoma to immunotherapy. *Nature Medicine*, 26(5):781–791, 2020.

[177] Benjamin I. Philipson, Roddy S. O'Connor, Michael J. May, Carl H. June, Steven M. Albelda, and Michael C. Milone. 4-1BB costimulation promotes CAR T cell survival through noncanonical NF-κB signaling. *Science Signaling*, 13(625), 2020.

[178] P. Priestley, J. Baber, M.P. Lolkema, N. Steeghs, E. Bruijn, and C. Shale. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, 575:210–216.

[179] W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, and P.W. Zandstra. Pert: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput Biol*, 8(12).

[180] Sergio A. Quezada, Tyler R. Simpson, Karl S. Peggs, Taha Merghoub, Jelena Vider, Xiaozhou Fan, Ronald Blasberg, Hideo Yagita, Pawel Muranski, Paul A. Antony, Nicholas P. Restifo, and James P. Allison. Tumor-reactive CD4+ T cells develop cytotoxic activity and eradicate large established melanoma after transfer into lymphopenic hosts. *Journal of Experimental Medicine*, 207(3):637–650, 2010.

[181] Gerald Quon, Syed Haider, Amit G. Deshwar, Ang Cui, Paul C. Boutros, and Quaid Morris. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine*, 5(3), 2013.

[182] Jüri Reimand, Juan M Vaquerizas, Annabel E Todd, Jaak Vilo, and Nicholas M Luscombe. Comprehensive reanalysis of transcription factor knockout expression data in saccharomyces cerevisiae reveals many new targets. *Nucleic acids research*, 38(14):4768–4777, 2010.

[183] Joseph M. Replogle, Thomas M. Norman, Albert Xu, Jeffrey A. Hussmann, Jin Chen, J. Zachery Cogan, Elliott J. Meer, Jessica M. Terry, Daniel P. Riordan, Niranjan Srinivas, Ian T. Fiddes, Joseph G. Arthur, Luigi J. Alvarado, Katherine A. Pfeiffer, Tarjei S. Mikkelsen, Jonathan S. Weissman, and Britt Adamson. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology*, 38(8):954–961, 2020.

[184] A. Reuben, C.N. Spencer, P.A. Prieto, V. Gopalakrishnan, S.M. Reddy, J.P. Miller, X. Mao, M.P. De Macedo, J. Chen, and X. Song. Genomic and immune heterogeneity are associated with differential responses to therapy in melanoma. NPJ genomic medicine 2.

[185] Nadeem Riaz, Jonathan J. Havel, Vladimir Makarov, Alexis Desrichard, Walter J. Urba, Jennifer S. Sims, F. Stephen Hodi, Salvador Martín-Algarra, Rajarsi Mandal, William H. Sharfman, Shailender Bhatia, Wen Jen Hwu, Thomas F. Gajewski, Craig L. Slingluff, Diego Chowell, Sviatoslav M. Kendall, Han Chang, Rachna Shah, Fengshen Kuo, Luc G.T. Morris, John William Sidhom, Jonathan P. Schneck, Christine E. Horak, Nils Weinhold, and Timothy A. Chan. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, 171(4):934–949.e15, 2017.

[186] A. Ribas and J.D. Wolchok. Cancer immunotherapy using checkpoint blockade. *Science*, 359:1350–1355.

[187] T. Ried. Homage to theodor boveri (1862-1915): Boveri's theory of cancer as a disease of the chromosomes, and the landscape of genomic imbalances in human carcinomas. *Environ. Mol. Mutagen*, 50:593–601.

[188] T. Ried, K. Heselmeyer-Haddad, H. Blegen, E. Schrock, and G. Auer. Genomic changes defining the genesis, progression, and malignancy potential in solid human tumors: a phenotype/genotype correlation. *Genes Chromosomes Cancer*, 25:195–204.

[189] T. Ried, Y. Hu, M.J. Difilippantonio, B.M. Ghadimi, M. Grade, and J. Camps. The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochim. Biophys. Acta*, 1819:784–793.

[190] T. Ried, R. Knutzen, R. Steinbeck, H. Blegen, E. Schrock, K. Heselmeyer, S. Manoir, and G. Auer. Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosomes Cancer*, 15:234–245.

[191] T. Ried, G.A. Meijer, D.J. Harrison, G. Grech, S. Franch-Exposito, R. Briffa, B. Carvalho, and J. Camps. The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. *Mol. Aspects Med*, 69:48–61.

[192] N.A. Rizvi, M.D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J.J. Havel, W. Lee, J. Yuan, P. Wong, and T.S. Ho. Cancer immunology. mutational landscape determines sensitivity to pd-1 blockade in non-small cell lung cancer. *Science*, 348:124–128.

[193] Paul F. Robbins, Yong Chen Lu, Mona El-Gamil, Yong F. Li, Colin Gross, Jared Gartner, Jimmy C. Lin, Jamie K. Teer, Paul Cliften, Eric Tycksen, Yardena Samuels, and Steven A. Rosenberg. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nature Medicine*, 19(6):747–752, 2013.

[194] Christelle Robert and Mick Watson. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, 16(1), 2015.

[195] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[196] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), 2010.

[197] Whijae Roh et. al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Science Translational Medicine*, 9(379), 2017.

[198] E. Rollman, M.Z. Smith, A.G. Brooks, D.F. Purcell, B. Zuber, I.A. Ramshaw, and S.J. Kent. Killing kinetics of simian immunodeficiency virus-specific cd8+ t cells: implications for hiv vaccine strategies. *Journal of immunology*, 179:4571–4579.

[199] M.S. Rooney, S.A. Shukla, C.J. Wu, G. Getz, and N. Hacohen. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, 160:48–61.

[200] L.M. Sack, T. Davoli, M.Z. Li, Y. Li, Q. Xu, and K. Naxerova. Profound tissue specificity in proliferation control underlies cancer drivers and aneuploidy patterns. *Cell*, 173:499–514.

[201] M. Sade-Feldman, K. Yizhak, S.L. Bjorgaard, J.P. Ray, C.G. Boer, R.W. Jenkins, D.J. Lieb, J.H. Chen, D.T. Frederick, and M. Barzily-Rokni. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175:998–1013 1020.

[202] U. Sahin and O. Tureci. Personalized vaccines for cancer immunotherapy. *Science*, 359:1355–1360.

[203] Sayed Mohammad Ebrahim Sahraeian, Marghoob Mohiyuddin, Robert Sebra, Hagen Tilgner, Pegah T. Afshar, Kin Fai Au, Narges Bani Asadi, Mark B. Gerstein, Wing Hung Wong, Michael P. Snyder, Eric Schadt, and Hugo Y.K. Lam. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, 8(1), 2017.

[204] R.M. Samstein, C.-H. Lee, A.N. Shoushtari, M.D. Hellmann, R. Shen, Y.Y. Janjigian, D.A. Barron, A. Zehir, E.J. Jordan, and A. Omuro. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature genetics*, 51:202–206.

[205] Martin H. Schaefer, Jean Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E. Wanker, and Miguel A. Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2), 2012.

[206] Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, 2006.

[207] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(88):1–13, 2007.

[208] P. Sharma, S. Hu-Lieskovan, J.A. Wargo, and A. Ribas. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*, 168:707–723.

[209] J.M. Sheltzer, J.H. Ko, J.M. Replogle, N.C. Habibe Burgos, E.S. Chung, C.M. Meehl, N.M. Sayles, V. Passerini, Z. Storchova, and A. Amon. Single-chromosome gains commonly function as tumor suppressors. *Cancer Cell*, 31:240–255.

[210] John Paul Shen, Dongxin Zhao, Roman Sasik, Jens Luebeck, Amanda Birmingham, Ana Bojorquez-Gomez, Katherine Licon, Kristin Klepper, Daniel Pekin, Alex N. Beckett, Kyle Salinas Sanchez, Alex Thomas, Chih Chung Kuo, Dan Du, Assen Roguev, Nathan E. Lewis, Aaron N. Chang, Jason F. Kreisberg, Nevan Krogan, Lei Qi, Trey Ideker, and Prashant Mali. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, 14(6):573–576, 2017.

[211] S.S. Shen-Orr, R. Tibshirani, and P. Khatri. Cell type-specific gene expression differences in complex tissues. *Nat Methods*, 7(4):287–289.

[212] Weiwei Shi, Charlotte K.Y. Ng, Raymond S. Lim, Tingting Jiang, Sushant Kumar, Xiaotong Li, Vikram B. Wali, Salvatore Piscuoglio, Mark B. Gerstein, Anees B. Chagpar, Britta Weigelt, Lajos Pusztai, Jorge S. Reis-Filho,

and Christos Hatzis. Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. *Cell Reports*, 25(6):1446–1457, 2018.

[213] Yael Silberberg, Martin Kupiec, and Roded Sharan. A method for predicting protein-protein interaction types. *PloS one*, 9(3):e90904, 2014.

[214] Dana Silverbush and Roded Sharan. Network orientation via shortest paths. *Bioinformatics*, 30(10):1449–1455, 2014.

[215] Lillian L. Siu, Neeltje Steeghs, Tarek Meniawy, Markus Joerger, Jennifer L. Spratlin, Sylvie Rottey, Adnan Nagrial, Adam Cooper, Roland Meier, Xiaowei Guan, Penny Phillips, Gaurav Bajaj, Jochem Gokemeijer, Alan J. Korman, Kyaw Lwin Aung, and Matteo S. Carlino. Preliminary results of a phase I/IIa study of BMS-986156 (glucocorticoid-induced tumor necrosis factor receptor–related gene [GITR] agonist), alone and in combination with nivolumab in pts with advanced solid tumors. *Journal of Clinical Oncology*, 35(15_suppl):104–104, 2017.

[216] Mark J. Smyth, Shin Foong Ngiow, Antoni Ribas, and Michele W.L. Teng. Combination cancer immunotherapies tailored to the tumour microenvironment, 2016.

[217] A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J.M. Zaretsky, A. Desrichard, L.A. Walsh, M.A. Postow, P. Wong, and T.S. Ho. Genetic basis for clinical response to ctla-4 blockade in melanoma. *The New England journal of medicine*, 371:2189–2199.

[218] S. Spranger. Tumor heterogeneity and tumor immunity: A chicken-and-egg problem. *Trends in immunology*, 37:349–351.

[219] E. Staub, A. Rosenthal, and B. Hinzmann. Systematic identification of immunoreceptor tyrosine-based inhibitory motifs in the human proteome. *Cell Signal*, 16(4):435–456.

[220] M. Steri, V. Orrù, and M.L. Idda. Overexpression of the cytokine baff and autoimmunity risk. *N Engl J Med*, 376(17):1615–1626.

[221] S. Stingele, G. Stoehr, K. Peplowska, J. Cox, M. Mann, and Z. Storchova. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol. Syst. Biol*, 8(608).

[222] E. Stronen, M. Toebes, S. Kelderman, M.M. Buuren, W. Yang, N. Rooij, M. Donia, M.L. Boschen, F. Lund-Johansen, and J. Olweus. Targeting of cancer neoantigens with donor-derived t cell receptor repertoires. *Science*, 352:1337–1341.

[223] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[224] V. Subbiah, D. B. Solit, T. A. Chan, and R. Kurzrock. The FDA approval of pembrolizumab for adult and pediatric patients with tumor mutational burden (TMB) 10: a decision centered on empowering patients and their physicians, 2020.

[225] Atsushi Tanaka and Shimon Sakaguchi. Regulatory T cells in cancer immunotherapy, 2017.

[226] A.M. Taylor, J. Shih, G. Ha, G.F. Gao, X. Zhang, and A.C. Berger. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell*, 33:676–689.

[227] Mingxiang Teng, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, Christopher E. Mason, Sara Olson, Dmitri Pervouchine, Cricket A. Sloan, Xintao Wei, Lijun Zhan, and Rafael A. Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17(1), 2016.

[228] A.E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29:189–196.

[229] Anthony W. Tolcher, Mario Sznol, Siwen Hu-Lieskovan, Kyriakos P. Papadopoulos, Amita Patnaik, Drew W. Rasco, Donna Di Gravio, Bo Huang, Dhiraj Gambhire, Ying Chen, Aron D. Thall, Nuzhat Pathan, Emmett V. Schmidt, and Laura Q.M. Chow. Phase Ib study of utomilumab (PF-05082566), a 4-1BB/CD137 agonist, in combination with pembrolizumab (MK-3475) in patients with advanced solid tumors. *Clinical Cancer Research*, 23(18):5349–5357, 2017.

[230] E. Tran, S. Turcotte, A. Gros, P.F. Robbins, Y.-C. Lu, M.E. Dudley, J.R. Wunderlich, R.P. Somerville, K. Hogan, and C.S. Hinrichs. Cancer immunotherapy based on mutation-specific cd4+ t cells in a patient with epithelial cancer. *Science*, 344:641–645.

[231] D. Tsafrir, M. Bacolod, Z. Selvanayagam, I. Tsafrir, J. Shia, and Z. Zeng. Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res*, 66:2129–2137.

[232] M.B. Upender, J.K. Habermann, L.M. McShane, E.L. Korn, J.C. Barrett, M.J. Difilippantonio, and T. Ried. Chromosome transfer induced aneuploidy results in complex dysregulation of the cellular transcriptome in immortalized and cancer cells. *Cancer Res*, 64:6941–6949.

[233] E.M. Van Allen, D. Miao, B. Schilling, S.A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M.H. Geukes Foppen, and S.M. Goldinger. Genomic

correlates of response to ctla-4 blockade in metastatic melanoma. *Science*, 350:207–211.

[234] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1), 2010.

[235] M.M. Varin, L. Le Pottier, P. Youinou, D. Saulep, F. Mackay, and J.O. Pers. B-cell tolerance breakdown in sjögren's syndrome: Focus on baff. *Autoimmun Rev*, 9(9):604–608.

[236] J.R. Veatch, S.M. Lee, M. Fitzgibbon, I.T. Chow, B. Jesernig, T. Schmitt, Y.Y. Kong, J. Kargl, A.M. Houghton, and J.A. Thompson. Tumor-infiltrating brafv600e-specific cd4+ t cells correlated with complete clinical response in melanoma. *The Journal of clinical investigation*, 128:1563–1568.

[237] Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799, 2004.

[238] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics, 2016.

[239] J. Wang, C.J. Perry, K. Meeth, D. Thakral, W. Damsky, G. Micevic, S. Kaech, K. Blenman, and M. Bosenberg. Uv-induced somatic mutations elicit a functional t cell response in the yummer1.7 mouse melanoma model. *Pigment cell melanoma research*, 30:428–435.

[240] Kun Wang, Sushant Patkar, Joo Sang Lee, E. Michael Gertz, Welles Robinson, Fiorella Schischlik, David R. Crawford, Alejandro A. Schäffer, and Eytan Ruppin. Deconvolving clinically relevant cellular immune crosstalk from bulk gene expression using codefacs and lirics. *bioRxiv*, 2021.

[241] N. Wang, E.P. Hoffman, and L. Chen. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci Rep*, 6.

[242] L.K. Ward-Kavanagh, W.W. Lin, Ware Šedý, JR, and C.F. The tnf receptor superfamily in co-stimulating and co-inhibitory responses. *Immunity*, 44(5):1005–1019.

[243] T.B.K. Watkins, E.L. Lim, M. Petkovic, S. Elizalde, N.J. Birkbak, and G.A. Wilson. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, 587:126–132.

[244] B.A. Weaver and D.W. Cleveland. The aneuploidy paradox in cell growth and tumorigenesis. *Cancer Cell*, 14:431–433.

[245] M.J. Williams, B. Werner, C.P. Barnes, T.A. Graham, and A. Sottoriva. Identification of neutral tumor evolution across cancer types. *Nat Genet*, 48:238–244.

[246] Egon Willighagen, Michel Ballings, and Maintainer Michel Ballings. Package 'genalg'. 2015.

[247] M. Wolf, S. Mousses, S. Hautaniemi, R. Karhu, P. Huusko, M. Allinen, A. Elkahloun, O. Monni, Y. Chen, A. Kallioniemi, and Kallioniemi O.P. High-resolution. analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. Neoplasia. 2004;6:240-247.

[248] Y. Wolf and Y. Samuels. Cancer research in the era of immunogenomics. *ESMO open*, 3:000475.

[249] Yochai Wolf, Osnat Bartok, Sushant Patkar, Gitit Bar Eli, Sapir Cohen, Kevin Litchfield, Ronen Levy, Alejandro Jiménez-Sánchez, Sophie Trabish, Joo Sang Lee, et al. Uvb-induced tumor heterogeneity diminishes immune response in melanoma. *Cell*, 179(1):219–235, 2019.

[250] Douglas C. Wu, Jun Yao, Kevin S. Ho, Alan M. Lambowitz, and Claus O. Wilke. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19(1), 2018.

[251] J.H. Xue and D.M. Titterington. Median-based image thresholding. *Image Vis Comput*, 29(9):631–637.

[252] Mark Yarchoan, Alexander Hopkins, and Elizabeth M. Jaffee. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *New England Journal of Medicine*, 377(25):2500–2501, 2017.

[253] Chen-Hsiang Yeang, Trey Ideker, and Tommi Jaakkola. Physical network models. *Journal of computational biology*, 11(2-3):243–262, 2004.

[254] Haiyuan Yu, Pascal Braun, Muhammed A Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[255] K. Zaitsev, M. Bambouskova, A. Swain, and M.N. Artyomov. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat Commun*, 10(1).

[256] M. Zanetti. Tapping cd4 t cells for cancer immunotherapy: the choice of personalized genomics. *Journal of immunology*, 194:2049–2056.

[257] Y. Zhong, Y.W. Wan, K. Pang, L.M.L. Chow, and Z. Liu. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, 14.